HAL
open science

Reconnaissance de mélange de densité et classification.Un algorithme d'apprentissage probabiliste : l'algorithme SEM

Gilles Celeux, Jean Diebolt

To cite this version:

Gilles Celeux, Jean Diebolt. Reconnaissance de mélange de densité et classification.Un algorithme d'apprentissage probabiliste : l'algorithme SEM. RR-0349, INRIA. 1984. inria-00076208

HAL Id: inria-00076208

https://inria.hal.science/inria-00076208

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

Rapports de Recherche

N° 349

RECONNAISSANCE DE MÉLANGE DE DENSITÉ ET CLASSIFICATION

UN ALGORITHME D'APPRENTISSAGE PROBABILISTE : L'ALGORITHME SEM

Gilles CELEUX
Jean DIEBOLT

Novembre 1984

RECONNAISSANCE DE MELANGE DE DENSITES ET CLASSIFICATION UN ALGORITHME D'APPRENTISSAGE PROBABILISTE: L'ALGORITHME SEM

Gilles CELEUX, Jean DIEBOLT

INRIA , PARIS 6

Résumé

Dans une première partie, nous faisons une revue commentée des principales méthodes de reconnaissance de mélanges de densités en distinguant l'approche classification et l'approche estimation.

Nous montrons les rapports existants entre les algorithmes usuels de classification et le problème de reconnaissance de mélanges de densité.

Nous présentons ensuite un algorithme d'apprentissage probabiliste (algorithme SEM) dérivé de l'algorithme EM pour ce problème.

Nous donnons les caractéristiques de l'algorithme SEM, étudions son comportement asymptotique et illustrons son intérêt à l'aide d'applications.

Nous donnons de plus deux versions séquentielles de cet algorithme et étudions leur comportement asymptotique.

Abstract

In a first part, we give a commented survey about estimation of mixture parameters. We distinguish the clustering approach and the estimation approach.

We show relations between usual clustering algorithms and the mixture problem.

Then we present a probabilistic teacher algorithm (SEM algorithm) derived from EM algorithm for this problem.

We discuss characteristics of SEM algorithm, study its asymptotical behaviour and present some applications.

Then we give two sequential versions of this algorithm and study their asymptotical behaviour.

PAPIER RÉCUPERE ET RECYCLE

RECONNAISSANCE DE MELANGES DE DENSITES ET CLASSIFICATION
UN ALGORITHME D'APPRENTISSAGE PROBABILISTE: L'ALGORITHME SEM

1. Introduction

1.1. Le problème traité

L'algorithme SEM (Stochastique, Estimation, Maximisation) a pour but de déterminer les composants d'un mélange fini de densités de lois de probabilité, ainsi que le nombre lui-même de ces composants, par une approche d'apprentissage probabiliste.

Le problème de reconnaissance de mélanges est le suivant:

Soit un échantillon $E=(x_1,\ldots,x_N)$ d'une variable aléatoire X à valeurs dans R^d, dont la loi admet la densité:

$$f(x)=\Sigma(p_k f(x,a_k);k=1,K) \quad \text{avec:}$$

$$\forall k=1,K \quad 0<p_k<1 \text{ et } \Sigma(p_k;k=1,K)=1$$

$f(x,a)$ est la densité de probabilité dépendant du paramètre vectoriel a de R^s.
p_k est la probabilité qu'un point de l'échantillon suive la loi de densité $f(x,a_k)$.

Le problème consiste à estimer le nombre K de composants et les paramètres inconnus
$(q_k=(p_k,a_k):k=1,K)$ au vu de l'échantillon .

Ce problème ne peut se résoudre, dans toute sa généralité, que si le mélange de densités appartient à une famille de mélanges identifiable.

Une famille F de mélanges est identifiable si et seulement:

$$f(.)\in F,\ f(.)=\Sigma(p_k f(.,a_k);k=1,K)=\Sigma(p'_k f(.,a'_k);k=1,K')$$

$$\Longrightarrow K=K' \quad \text{et} \quad \forall k=1,K\ p_k=p'_k,\ a_k=a'_k$$

Teicher <Te63>, Yakowitz et Spragins <Ya,Sp68> ont donné des théorèmes de caractérisation des mélanges identifiables dont le principal est le suivant:

Théorème: Une condition nécessaire et suffisante pour qu'une famille F de mélanges soit identifiable est que la famille

$\{F(.,a), a \in R^s\}$ soit un système libre sur R.

Ces auteurs ont ainsi montré que les mélanges gaussiens, exponentiels , de Poisson, de Cauchy sont identifiables. Par contre, les mélanges binomiaux et uniformes ne le sont en général pas.

1.2. Motivations de ce travail

En classification automatique et en reconnaissance des formes, les algorithmes existants sont nombreux. Il apparait de plus en plus qu'il faut procéder à une exploration fine des domaines de validité de tel ou tel algorithme.

Une manière possible de procéder consiste à essayer un tel algorithme sur des données simulées, engendrées selon un mécanisme probabiliste.

Un mécanisme de cette sorte est fourni par le tirage aléatoire d'un échantillon selon une loi de probabilité mélange de K densités de probabilité comme décrit ci-dessus.

C'est dans ce cadre que nous avons voulu:

(a) Préciser le domaine de validité des méthodes de partitionnement fondées sur la minimisation de l'inertie intra-classe: Méthodes de type nuées dynamiques (cf <Di80>), algorithme d'échange,...

(b) Résoudre le problème, encore très largement ouvert, de la détermination du nombre de classes.

(c) Tenir compte, dans la forme même des résultats fournis du degré de fiabilité de ces résultats; laquelle dépend en particulier de l'imbrication des composants du mélange.

(d) Détecter lorsque cela a lieu l'existence de plusieurs composants de moyennes très voisines (voire identiques) mais caractérisés par des structures différentes de la dispersion.

Pour y parvenir, nous avons été amené à concevoir un nouvel algorithme (algorithme SEM) de reconnaissance de mélange de densités de probabilité.

1.3. Plan du rapport

Au paragraphe 2, nous faisons une revue bibliographique commentée des principales méthodes existantes.

Certaines de ces méthodes se placent dans l'optique de la classification ce qui constitue une démarche qui ne vise pas la même finalité que celle exposée ci-dessus (cf paragraphe 1.2).

L'exposé de ces méthodes nous amène à préciser le lien existant entre les algorithmes usuels de classification fondés sur la minimisation de l'inertie intra-classe et le problème de reconnaissance de mélanges gaussiens (cf paragraphe 2.1.3).

Au paragraphe 3, nous nous intéressons ensuite aux méthodes d'inspiration statistique. Il en ressort que l'algorithme le plus efficace est l'algorithme EM qui recherche à maximiser la vraisemblance de l'échantillon relativement au modèle de mélange par un procédé itératif faisant intervenir deux étapes d'inférence statistique (estimation et maximisation).

Nous étudions les limites de l'algorithme EM:

-Il exige de connaitre le nombre exact de composants du mélange.

-Les résultats obtenus dépendent de l'initialisation:

-Il peut converger vers un point stationnaire de la vraisemblance de type col.

-Il peut converger avec une lenteur rédhibitoire.

Du point de vue théorique, les nombreux travaux qui lui ont été consacrés n'ont fourni des preuves de convergence que sous des hypothèses difficiles à vérifier en pratique (cf paragraphe 2.2.2).

Au paragraphe 3, nous présentons l'algorithme SEM qui s'inspire de l'algorithme EM en lui adjoignant une étape d'apprentissage probabiliste.

Nous précisons les caractéristiques de cet algorithme et la forme des résultats qu'il obtient.

Le paragraphe 4 est consacré à l'étude théorique de l'algorithme SEM.

Dans le cas simple d'un mélange à deux composants où seul le paramètre de proportion p est inconnu, nous étudions le comportement asymptotique lorsque la taille N de l'échantillon tend vers l'infini de

la loi stationnaire de la suite des itérés de l'algorithme SEM (théorème 2). Nous utilisons pour cela un résultat (théorème 1) concernant une chaine de Markov constituant une modélisation approximative du fonctionnement de l'algorithme SEM.

Ces deux théorèmes sont des résultats mathématiques nouveaux.

Nous conjecturons que ces résultats peuvent s'étendre au cas général.

Au paragraphe 5 nous présentons des utilisations de l'algorithme SEM sur des données simulées.

Au paragraphe 6 nous présentons des utilisations de l'algorithme SEM sur des données réelles.

Aux paragraphes 7 et 8 nous présentons deux versions séquentielles de l'algorithme SEM et étudions leurs comportements asymptotiques dans le même cas simple qu'au paragraphe 4.

Dans ce cadre, la version séquentielle stricte présenté au paragraphe 7, s'avère très proche d'un algorithme d'apprentissage probabiliste de type approximation stochastique étudié par Silverman (cf <Si80>).

La version séquentielle large présenté au paragraphe 8, qui ne relève pas des méthodes de l'approximation stochastique, donne lieu à deux résultats mathématiques nouveaux analogues aux théorèmes 1 et 2 du paragraphe 4: les théorèmes 3 et 4.

2. Les deux approches du problème

Le problème de reconnaissance de mélanges a été étudié par de nombreux auteurs sous des hypothèses plus ou moins restrictives.

Dans les deux paragraphes suivants nous faisons une revue commentée des principales méthodes de reconnaissance de mélanges en distinguant deux approches foncièrement différentes. Dans cette revue, nous insistons sur les méthodes les plus reliées à notre approche que nous présentons ensuite.

2.1. Approche classification

2.1.1. Présentation

Cette approche (<ScSy71>,<Sch76>,<Sy81>) consiste à rechercher une partition

$P=(P_1,\ldots,P_K)$

telle que chaque classe P_k soit assimilable à un sous échantillon suivant la loi $f(.,a_k)$.

Dans ce cadre, les algorithmes utilisés sont de type Nuées Dynamiques (<Di80>). Aussi nous présentons l'algorithme d'A.Schroeder (<Sch76>), valide pour toute famille paramétrée de densités identifiable dont les paramètres admettent des estimateurs du maximum de vraisemblance.

L'espace de représentation d'une partition étant l'espace des paramètres, la méthode vise à maximiser le critère de vraisemblance classifiante suivant, qui mesure l'adéquation d'une partition et de sa représentation:

$$W(a,P)= \Sigma(\text{Log } L(P_k,a_k);k=1,K)$$

où la fonction $L(P_k,a_k)$

est la vraisemblance du sous échantillon P_k pour la loi de densité $f(x,a_k)$.

L'algorithme se déroule ainsi: à partir d'une partition P^o en K classes de l'échantillon, on applique successivement deux fonctions g et h jusqu'à obtention d'une partition stable.

g, appellée fonction de représentation, est définie par:

$g(P)=g(P_1,\ldots,P_k)=(a_1,\ldots,a_k)$ avec

a_k : estimation du maximum de vraisemblance du paramètre de la densité associée au sous échantillon P_k.

h, appellée fonction d'affectation, est définie par:

$h(a)=h(a_1,\ldots,a_k)=(P_1,\ldots,P_k)$ avec

$$P_k=\{x/f(x,a_k)\geqslant f(x,a_m),\text{ avec } k<m \text{ en cas d'égalité}\}$$

Cet algorithme fait croitre le critère en un nombre fini d'itérations. A la convergence, on obtient une partition P et une estimation des paramètres $(a_1,..,a_k)$.

Les paramètres $(p_k;k=1,K)$ sont estimés par les quantités $(\mathrm{card}P_k/N;k=1,K)$.

2.1.2. Propriétés et limites

Cette approche présente avant tout l'intérêt d'être rapide.

Du point de vue théorique, la méthode peut être vue ainsi: Chaque point est issu de l'un des K composants du mélange. Considérons les N paramètres inconnus suivants $(o(i);i=1,N)$ avec:

$o(i)$=numéro du composant dont x_i est issu.

La méthode revient à rechercher les estimateurs du maximum de vraisemblance de ces paramètres. Or le nombre de paramètres à estimer augmente indéfiniment avec la taille N de l'échantillon. En conséquence les estimateurs du maximum de vraisemblance de ces paramètres ne sont pas convergents (<BrWy78>).

De plus, l'approche classification induit, en général, un biais dans l'estimation des paramètres du fait de la connexité des classes (<Ma75>).

Remarque:
Afin de réduire ce biais, des tentatives d'utilisation de cette méthode ont été faites en utilisant les probabilités a posteriori

$(t_k(x_i);k=1,K;i=1,N)$

d'appartenance des points de l'échantillon aux composants du mélange (<Ca76>,<Di80>).

En fait, l'étape de calcul des $t_k(x_i)$, les a_k étant fixés, conduit à prendre $\forall i=1,N\ t_k(x_i)=1$ ou 0.

En effet le critère à maximiser s'écrit alors:

$W(a,t)= \Sigma\ (t_k(x_i)\mathrm{Log}(f(x_i,a_k);i=1,N;k=1,K)$

L'étape de calcul des $t_k(x_i)$ consiste à maximiser $\forall i=1,N$ l'expression

$\Sigma(t_k(x_i)\mathrm{Log}(f(x_i,a_k);k=1,K)$

Or $\Sigma(t_k(x_i)\mathrm{Log}(f(x_i,a_k);k\neq 1,K)\leqslant \mathrm{Log}(f(x_i,a_m))$

avec m vérifiant $\forall k\neq m\ f(x_i,a_k)\leqslant f(x_i,a_m)$

La borne de $\Sigma(t_k(x_i)\mathrm{Log}(f(x_i,a_k);k=1,K)$ est donc atteinte pour $t_k(x_i)=1$ si $k=m$, 0 sinon.

2.1.3. Etude comparative dans le cas gaussien avec les algorithmes usuels de partitionnement

Dans ce paragraphe, nous allons voir que les algorithmes de partitionnement utilisant des critères d'inertie peuvent se présenter comme des méthodes pour résoudre le problème de reconnaissance de mélanges gaussiens dans l'optique classification (<ScSy71>,<BCD81>). Dans le cas d'un mélange gaussien, les paramètres $(a_k;k=1,K)$ s'écrivent $a_k=(m_k,\Gamma_k)$ avec:

m_k=espérance du composant numéro k.

Γ_k=matrice variance du composant numéro k.

Le critère de vraisemblance classifiante s'écrit:

$W(a,P)=\mathrm{Cste}-1/2\ \Sigma(\ \Sigma((x_i-m_k)'\ \Gamma_k^{-1}(x_i-m_k)+\mathrm{Log}\ |\Gamma_k|;x_i\epsilon P_k);k=1,K)$

Donc maximiser W(a,P) revient à minimiser le critère:

$C(a,P)= \Sigma(\ \Sigma((x_i-m_k)'\ \Gamma_k^{-1}(x_i-m_k)+\mathrm{Log}\ |\Gamma_k|;x_i\ \epsilon\ P_k);k=1,K)$

que nous considérons dans la suite.

<u>Premier cas:</u> $\Gamma_k=\Gamma$, $\forall k=1,K$; Γ supposé connu.

Le critère à minimiser devient:

$$C(a,P)=\Sigma(\Sigma((x_i-\hat{m}_k)'\Gamma_k^{-1}(x_i-\hat{m}_k);x_i\in P_k);k=1,K)$$

avec $\hat{m}_k$: moyenne empirique des élements de la classe P_k.

Γ étant une matrice symétrique, il existe une matrice T telle que $\Gamma=TT'$.

Par le changement de variable $y=T^{-1}x$, on peut se ramener au cas où $\Gamma=I$. Donc le critère peut s'écrire:

$$C(a,P)=\Sigma(\Sigma((x_i-\hat{m}_k)'(x_i-\hat{m}_k);x_i\in P_k);k=1,K)$$

Il s'agit du critère de la version la plus simple et la plus utilisée des Nuées Dynamiques (algorithme des "centres mobiles").

Ainsi s'explique le fait, constaté expérimentalement, que cette méthode ait tendance à donner des classes sphériques de même volume.

Deuxième cas: $\Gamma_k=\Gamma$, $\forall k=1,K$; Γ supposé inconnu.

Dans ce cas, la partition étant fixée, l'estimateur du maximum de vraisemblance de Γ est W/N où:

$$W=\Sigma(\Sigma((x_i-\hat{m}_k)(x_i-\hat{m}_k)';x_i\in P_k);k=1,K)$$

Le critère à minimiser s'écrit:

$$C(a,P)=\Sigma(\Sigma(N^{1/d}(x_i-\hat{m}_k)'W^{-1}(x_i-\hat{m}_k)+\mathrm{Log}(|W/N|);x_i\in P_k);k=1,K)$$

$$C(a,P)=N^{1/d}\mathrm{tr}(WW^{-1})+\Sigma(\mathrm{card}P_k\mathrm{Log}|W|;k=1,K)+\mathrm{Cste}$$

Il se réduit donc à:

$$C(a,P)=|W|$$

Ce critère a été proposé, sans aucune référence au modèle gaussien dans <FrRu67> et <Go75> dans le but de reconnaitre des classes ayant le même type de dispersion mais possédant des directions d'allongement inconnues.

Troisième cas: les Γ_k sont différents entre eux et inconnus

A chaque étape, la partition étant fixée, les estimateurs du maximum de vraisemblance des Γ_k sont $V_k=W_k/\mathrm{card}P_k$ avec:

$$W_k=(\Sigma((x_i-\hat{m}_k)(x_i-\hat{m}_k)';x_i\in P_k)$$

Le critère à minimiser s'écrit:

$$C(a,P)=\Sigma(\Sigma(N^{1/d}(x_i-\hat{m}_k)'V_k^{-1}(x_i-\hat{m}_k)+\mathrm{card}P_k\mathrm{Log}(|V_k|);x_i\in P_k);k=1,K)$$

$$C(a,P)=\Sigma(tr(cardP_k V_k V_k^{-1})+cardP_k Log|V_k|\ ;k=1,K)$$

et se réduit à:

$$C(a,P)=\Sigma(cardP_k Log|V_k|;k=1,K)$$

Par ailleurs, le critère de l'algorithme des distances adaptatives (<Go75>) élaboré pour permettre de reconnaitre des classes de "formes" différentes est:

$$C'(a,P)=\Sigma(\Sigma((x_i-\hat{m}_k)'|V_k|^{1/d}V_k^{-1}(x_i-\hat{m}_k);x_i \in P_k);k=1,K)$$

$$C'(a,P)=\Sigma(cardP_k|V_k|^{1/d}trV_k V_k^{-1};k=1,K)$$

il devient:

$$C'(a,P)=\Sigma(cardP_k|V_k|^{1/d};k=1,K)$$

Les deux critères ne sont pas les mêmes, mais sont très analogues. En pratique, les deux méthodes donnent des résultats quasi identiques (<Go75>).

2.2. Approche estimation

Cette approche qui est l'approche directe est la plus ancienne et la plus répandue. Les principales techniques d'estimation utilisées sont celles des moments (<Pe94>,<Co64>,<QuRa78>...) et du maximum de vraisemblance (<Sh68>,<Da69>, <Wo70>,<Ka77>,...).

2.2.1. Méthode des moments

Le principe de cette méthode consiste à résoudre les équations:

$$\int(x-E(x))^s f(x)dx=1/N\ \Sigma((x_i-x)^s;i=1,N)$$

s variant de 1 au nombre de moments nécessaires pour estimer les paramètres du mélange.

Ce type de méthode présente l'avantage de ne pas donner de solutions singulières (voir paragraphe 2.2.3).

Malheureusement la méthode des moments n'est pas praticable pour un grand nombre de composants ou pour des distributions multidimensionnelles. Les raisons de cette limitation sont algébriques (les équations deviennent inextricables) et statistiques (la variance

des estimateurs des moments d'ordre plus grand ou égal à deux devient très grande lorsque N augmente).

Dans le cas d'un mélange gaussien à deux composants Quandt et Ramsey (<QuRa78>) ont proposé une méthode qui utilise les moments de la fonction génératrice du mélange E(exp(cx)).

Pour une valeur donnée de c, ils considèrent l'estimation:

$\tilde{E}(\exp(c_j x))=1/N\,\Sigma(\exp(c_j x_i);i=1,N)$

Les paramètres du mélange sont estimés par minimisation de

$S=\Sigma((\tilde{E}(\exp(c_j x)-E(\exp(c_j x))^2;j=1,5)$

par une méthode des moindres carrés.

Ils obtiennent ainsi des estimateurs convergents des paramètres. Les résultats sont meilleurs que ceux obtenus par la méthode des moments si les paramètres $(c_j;j=1,5)$ de la méthode sont bien choisis.

2.2.2. Methode du maximum de vraisemblance

Cette méthode consiste à résoudre itérativement les équations de vraisemblance; le Log de la vraisemblance étant:

$L(x_1,..,x_N,a_1,..,a_k,p_1,..,p_k)=\Sigma(\mathrm{Log}(\Sigma(p_k f(x_i,a_k);k=1,K));i=1,N)$

Les algorithmes les plus efficaces pour résoudre les équations de vraisemblance sont, à des variantes près, des algorithmes de type EM (<DLR77>) (cf <Sh69>,<Wo70>,<ReWa84>...).

2.2.2.1 L'algorithme EM

A partir d'une solution initiale $(p_k^o,a_k^o;k=1,K)$, l'algorithme est le suivant:

étape E (estimation):

Pour k=1,K;i=1,N

calcul des $t_k^n(x_i)=p_k^n f(x_i,a_k^n)/\Sigma(p_k^n f(x_i,a_k^n);k=1,K)$

étape M (maximisation):

Pour k=1,K calcul de $p_k^{n+1}=1/N\ \Sigma(t_k^n(x_i);i=1,N)$

et résolution des équations pour k=1,K;j=1,s:

$$\Sigma(t_k^n(x_i)\partial\log f(x_i,a_k^{n+1})/\partial a_{jk},i=1,N)=0$$

où $a_k=(a_{jk};j=1,s)$

L'algorithme EM présente les caractéristiques suivantes:

-Il fonctionne pour un grand nombre de composants et dans le cas multidimensionnel.
-Il fournit, en général, de bons résultats si le nombre de composants est connu.
-Malheureusement, il converge extrêmement lentement. Cette lenteur peut rendre son utilisation rédhibitoire. C'est en particulier le cas lorsque la solution initiale est éloignée de la solution limite accessible.

Un problème de l'approche par le maximum de vraisemblance vient de ce que la fonction de vraisemblance n'est pas bornée. En conséquence, il se peut que l'algorithme EM dégénère vers une solution singulière (cf <EvHa81>). Cette possibilité peut être évitée en imposant des contraintes sur les paramètres du mélange. Par exemple Day (<Da69>) suppose l'égalité des matrices variance dans le cas d'un mélange gaussien. Malheureusement ce type de restriction est assez fort.

2.2.2.2 Résultats sur le comportement asymptotique de l'algorithme EM

Dans le cadre général de l'algorithme EM, qui dépasse celui des mélanges, et sous de larges hypothèses, Wu (<Wu83>) a montré que les points limites de toute suite des itérés engendrés par l'algorithme EM étaient des points stationnaires de la vraisemblance et qu'il existait un point fixe de l'algorithme atteignant cette valeur limite de la vraisemblance.

Boyles (<Bo83>) a montré un résultat analogue sous des hypothèses plus restrictives et plus difficiles à mettre en évidence.

Dans le cadre de la reconnaissance de mélange de densités, Redner et Walker (<ReWa84>) ont montré (en s'appuyant sur les résultats de Wu) que, si les densités appartenaient à la famille exponentielle, les points limites de toute suite des itérés engendrés par l'algorithme EM appartenaient à un ensemble compact et connexe constitué des points stationnaires de la vraisemblance et que tout point atteignant une

telle valeur limite de la vraisemblance était un point fixe de l'algorithme.

Ils ont montré, de plus, que si la matrice d'information de Fisher est définie positive pour les vraies valeurs des paramètres du mélange alors:

-Pour N assez grand, l'unique solution convergente q^N des équations de vraisemblance existe p.s.
-Il existe une norme sur l'espace des paramètres pour laquelle la suite des itérés de l'algorithme EM converge vers q^N si la solution initiale q^o est suffisamment proche de q^N.

Notons que ces résultats sont d'une grande portée car la quasi totalité des densités considérées dans les problèmes de mélanges appartiennent à la famille exponentielle.

2.2.3. Comparaison des méthodes des moments et du maximum de vraisemblance

L'approche par le maximum de vraisemblance s'avère supérieure à la méthode des moments pour les raisons suivantes:

-La méthode des moments est impraticable dans le cas multidimensionnel ou lorsque le nombre de composants est élevé.
-Des simulations ont montré que lorsque les composants du mélange étaient peu séparés les résultats par le maximum de vraisemblance étaient meilleurs. La méthode des moments n'obtient des résultats comparables que dans le cas gaussien unidimensionnel avec deux composants bien séparés.

Les résultats obtenus par la méthode des moments de la fonction génératrice sont meilleurs que ceux obtenus par la méthode des moments. Mais cette méthode reste limitée au cas unidimentionnel avec peu de composants. Par contre, elle présente l'intérêt de donner de bons résultats avec des échantillons de petite taille.

D'autre part, les auteurs de cette méthode récusent l'estimation par le maximum de vraisemblance du fait de la possibilité d'obtenir des solutions singulières.

On doit faire les remarques suivantes:

En pratique l'apparition de solutions singulières par l'algorithme EM est extrêmement rare. Ces solutions interviennent avant tout lorsque l'un des composants à une probabilité d'apparition p petite. Plus précisément les singularités ont lieu lorsque pour ce composant $Np<d$ (d étant la dimension de l'espace contenant

l'échantillon) ou plus généralement lorsque les points engendrés par l'un des composants sont tous situés dans un sous espace de codimension strictement positive.

Dans ce cas, il parait difficile d'obtenir une bonne estimation des paramètres de ce composant quelle que soit la méthode employée.

De plus des simulations ont montré (<Ho78>,<EvHa81>) que les résultats obtenus par l'algorithme EM étaient supérieurs à ceux obtenus par la méthode des moments de la fonction génératrice. En particulier cette méthode est sensible au choix de la solution initiale et au choix de paramètres qui lui sont inhérents (cf paragraphe 2.2.1)

2.2.4. Méthodes quasi bayésiennes et d'apprentissage probabiliste

Dans le cas où seules les proportions du mélange $p=(p_k,k=1,K)$ sont inconnues, une approche bayésienne du problème a été proposée par Smith et Makov (<SmMa78>).

Partant d'une distribution a priori $Pr^o(p)$ pour p, l'approche consiste à résoudre de manière séquentielle la formule de Bayes:

$$Pr(p/x_n)=f(x_n/p)Pr(p/x_{n-1})/\int f(x_n/p)Pr(p/x_{n-1})dp$$

Malheureusement, il n'existe pas de statistique exhaustive pour p et les calculs deviennent inextricables.

Les auteurs proposent une approximation de la formule itérative en partant d'une distribution de Dirichlet pour $Pr^o(p)$. Leur procédure converge p.s vers la vraie valeur du paramètre.

Cette approche peut également s'appliquer lorsque seuls les paramètres $(a_k,k=1,K)$ sont inconnus (<MaSm77>). Mais, dans ce cas, il n'existe pas de résultats de convergence, actuellement.

Une autre voie pour contourner la difficulté de résolution de la formule de Bayes consiste àà utiliser un algorithme d'apprentissage probabiliste.

Agrawala (<Ag70>) a construit un algorithme séquentiel pour estimer une valeur a parmi les a_k, les autres étant connues ainsi que les proportions du mélange.

Partant d'une distribution a priori $p^o(a)$ pour a, il procède ainsi:

x_n étant un nouveau point observé de l'échantillon, il l'affecte à l'un des composants par tirage aléatoire suivant la loi a posteriori

d'appartenance de x_n aux composants. Soit l_n le numéro du composant obtenu par ce tirage aléatoire.

Il recalcule une nouvelle probabilité a posteriori pour a, sous l'hypothèse que x_n est effectivement issu du composant l_n.

Là encore, il n'existe pas de résultats de convergence pour cet algorithme.

Par un procédé analogue, Silverman (<Si80>) traite le problème d'estimation de p_1 pour un mélange à deux composants dont les paramètres a sont connus.

Partant d'une loi a priori $Pr^o(p_1)=\beta(b_o,c_o)$, il procède ainsi:

Un nouveau point observé x_n de l'échantillon est affecté au composant 1 ou 2 par tirage aléatoire suivant sa loi a posteriori d'appartenance à ces composants.

Si x_n est affecté au composant 1 alors $b_n=b_{n-1}+1$.

Si x_n est affecté au composant 2 alors $c_n=c_{n-1}+1$.

La loi a posteriori de p_1 est une loi $\beta(b_n,c_n)$ et le rapport

$p_n=b_n/(b_n+c_n)$

est un estimateur de p_1.

Silverman montre que la suite des p_n converge p.s vers p_1.

En pratique, des simulations semblent montrer que le schéma d'apprentissage probabiliste donne des résultats meilleurs que le schéma quasi bayésien (<MaSm76>).

3. Notre approche: un algorithme d'apprentissage probabiliste (algorithme SEM)

Tous les algorithmes évoqués ci-dessus présentent les limitations suivantes:

- Le nombre K de composants est supposé connu.
- La solution obtenue dépend de la position initiale de l'algorithme.

L'algorithme que nous proposons ici répond en grande partie à ces deux limitations. Il utilise de manière complémentaire les idées des algorithmes de classification évoqués au paragraphe 2.1 et l'algorithme EM.

Il s'agit en fait d'un algorithme EM auquel nous avons adjoint une étape d'apprentissage probabiliste. D'où son nom, algorithme SEM: Stochastique, Estimation, Maximisation.

3.1. Présentation

Au départ, on fixe le paramètre K majorant supposé du nombre de composants du mélange et on se donne un seuil c(N) dépendant de la taille N de l'échantillon et de la dimension d. On se donne également une loi de probabilité sur l'espace des paramètres du mélange.

Initialisation:

En chaque point $x_i, i=1,N$ on choisit (en général au hasard) les probabilités initiales d'appartenance à l'un des composants:

Soient $t_k^o(x_i), k=1,K$ avec:

$0<t_k^o(x_i)<1$ et $\Sigma(t_k^o(x_i);k=1,K)=1$

Itération n (n>0):

Etape S (stochastique):

On tire en chaque point x_i la v.a multinomiale

$e^n(x_i)=(e_k^n(x_i);k=1,K)$

d'ordre un et de paramètres $(t_k^n(x_i);k=1,K)$

Les réalisations $e^n(x_i)$ définissent une partition $P^n=(P_1^n,\ldots,P_k^n)$ de l'échantillon avec:

$P_k^n=\{x_i/e_k^n(x_i)=1\}$

Si $card(P_k^n)$ est plus petit que $Nc(N)$, le phénomène est compatibilisé et l'on procède à un tirage des paramètres du mélange selon la loi donné à l'avance à partir desquels on procède à l'étape E décrite ci-dessous.

Sinon:

Etape M (maximisation):

On calcule les estimations du maximum de vraisemblance

$q_k^{n+1}=(p_k^{n+1},a_k^{n+1})$

des paramètres du mélange sur la base des sous échantillons $(P_k^n,k=1,K)$

on a:

$$p_k^{n+1}=1/N\ \Sigma(e_k^n(x_i);i=1,N)$$

L'estimation des a_k^{n+1} dépend bien sûr de la famille paramétrée, posée a priori, des composants du mélange.

Remarque: Si les paramètres $(a_k,k=1,K)$ n'admettent pas d'estimateur du maximum de vraisemblance, on calculera des estimateurs qui améliorent la vraisemblance comme il est fait dans <Sch76> pour un mélange de lois gamma ou plus généralement dans l'algorithme GEM (<DLR78>).

Dans le cas où l'espérance m_k ou la matrice de variance Γ_k sont des constituants des paramètres (cas de mélanges gaussiens, de Poisson, d'exponentielles,..) les estimations à l'itération n sont les suivantes:

$$m_k^{n+1}=\Sigma(e_k^n(x_i)x_i;i=1,N)/\Sigma(e_k^n(x_i);i=1,N)$$

$$\Gamma_k^{n+1}=\Sigma(e_k^n(x_i)(x_i-m_k^{n+1})(x_i-m_k^{n+1})';i=1,N)/\Sigma(e_k^n(x_i);1,N)$$

Etape E (estimation):

A partir des $q_k^{n+1}=(p_k^{n+1},a_k^{n+1})$,on calcule:

Pour $k=1,K;i=1,N$

$$t_k^{n+1}(x_i)=p_k^{n+1}f(x_i,a_k^{n+1})/\Sigma\,(p_k^{n+1}f(x_i,a_k^{n+1});k=1,K)$$

En pratique, dans les applications présentées aux paragraphes 5 et 6, nous avons simplifié la procédure.

-D'une part, on a pris $c(N)=d/N$.

-Lorsque pour un certainn k on a

$\mathrm{card}(P_k^n)\leqslant Nc(N)=d$

on supprime le composant numéro k et l'algorithme continue sur la base de (K-1) composants.

3.2. Caractéristiques de cette approche

A chaque itération l'algorithme construit une partition en classes en général non connexes représentatives des paramètres du mélange.

A la stabilité de l'algorithme, on obtient non pas une seule partition mais une suite de partitions statistiquement admissibles pour les estimations des paramètres du mélange. Ces estimations sont précises du fait de la non connexité des classes (cf les simulations du paragraphe 5) et asymptotiquement sans biais (cf paragraphe 4).

Le type de convergence obtenue est une convergence en loi correspondant à la stationnarité de la suite des estimés (p^n,a^n).

Par ailleurs, les perturbations introduites à chaque itération par les tirages aléatoires empêchent la convergence vers un maximum local instable de la vraisemblance comme cela peut être le cas avec l'algorithme EM.

En particulier, cet algorithme n'a pas besoin de connaitre le nombre de composants, mais seuleument un majorant de ce nombre. Ce point sera illustré aux paragraphes 5 et 6.

Enfin il converge notablement plus rapidement que l'algorithme EM quelle que soit la configuration initiale. Les tirages aléatoires l'empêchent de "stationner" trop longtemps loin de la solution limite.

3.3. Forme des résultats

Un argument classique de la théorie des chaines de Markov permet de montrer que la suite des variables aléatoires q^n converge en loi vers une mesure de probabilité, invariante sous la transition, définie sur l'espace des états définis ci-dessous:

Un état e est la donnée, pour chaque i, de l'affectation du point x_i à l'un des K composants. On a donc:

$e=(e_k(x_i),k=1,K;i=1,N)$

L'ensemble E des états comporte K^N éléments.

Du point de vue probabiliste, l'algorithme engendre une chaine de Markov homogène $(C_n,n\in N)$ à valeurs dans l'ensemble fini E, telle que:

$P(C_{n+1}=e/C_n=e')=$ probabilité de tirer e suivant la loi

$t(e')=(t_k(x_i;e');k=1,K;i=1,N)$

issue de e' par un calcul indépendant de n qui montre aussi que les $t_k(x_i;e')$ sont strictement positifs.

La chaine est donc apparemment irréductible, mais il existe des états absorbants: ceux pour lesquels une classe représentative d'un des composants au moins, comporte un nombre d'éléments trop faible pour permettre l'estimation des paramètres a_k, ou comporte des liaisons linéaires entre ses éléments. C'est l'apparition d'une solution singulière (cf paragraphe 2.2.3). Pour rendre la chaine irréductible il suffit, dans un tel cas, de relancer l'algorithme en effectuant à nouveau un tirage au hasard suivant une loi fixée d'avance.

La chaine C_n ainsi modifiée n'a qu'un seul état ergodique et la loi de C_n converge, à vitesse exponentielle, vers l'unique probabilité invariante sous la transition obtenue (cf <KeSn74>).

D'autre part, puisque le vecteur q^n se déduit de C_n par un calcul indépendant de n, il en résulte que la loi de q^n converge, vers la loi image de la loi stationnaire de la chaine $(C_n,n\in N)$.

La suite $(q^n,n\in N)$ est elle même une chaine de Markov homogène. Soit $\mathcal{E}=\Pi E_n$, les E_n étant des exemplaires de l'ensemble E, le fonctionnement de l'algorithme se traduit par une équation du type:

$q^{n+1}(e^o,\ldots,e^n)=T_N(q^n)+V_N(q^n,e^n)$

la variable aléatoire $V_N(q^n,e^n)$ étant indépendante de $T_N(q^n)$ conditionnellement à q^n.

En pratique, après que le nombre de classes se soit stabilisé, on édite à la stationnarité la moyenne et la variance de chacune des lois marginales de q^n.

Nous ne disposons pas actuellement de procédure statistique pour déterminer le nombre minimum d'itérations à partir duquel on peut considérer que la suite des paramètres

$q^n=(p^n,a^n)$

acquiert son comportement stationnaire (test de début d'enregistrement des résultats pour le calcul de la moyenne et de l'écart-type des lois marginales des q^n).

Pour l'instant, nous faisons tourner l'algorithme suffisament longtemps pour être assuré d'avoir atteint l'état stationnaire (phase d'apprentissage). Nous faisons ensuite tourner l'algorithme à partir de cet état stationnaire et nous enregistrons les résultats (phase d'exploitation).

4. Etude du comportement asymptotique de l'algorithme SEM dans un cas simple

4.1. Résultats préliminaires

On considére le cas d'un mélange à deux composants où seul le paramètre p^* $(0<p^*<1)$ est inconnu. On note:

$A_i=f_1(x_i), B_i=f_2(x_i)$

$t(x)=t_1(x)$ et $e(x)=e_1(x)$

Soit $J_N=[c(N),1-c(N)]$ (N étant la taille de l'échantillon).

L'algorithme s'écrit:

Si $p_N^n \in J_N$ est défini, on forme $p_N^{n+1/2}$:

$p_N^{n+1/2}=T_N(p_N^n)+V_N(p_N^n,e^n)$

avec:

$T_N(p^n)=1/N\,\Sigma(t^n(x_i);i=1,N)$ et

$V_N(p^n,e^n)=1/N\,\Sigma(e^n(x_i)-p^n(x_i);i=1,N)$.

Puis,

-Si $p_N^{n+1/2} \in J_N$, alors $p_N^{n+1}=p_N^{n+1/2}$

-Si $p_N^{n+1/2} \notin J_N$, alors p_N^{n+1}

est tiré au hasard selon une loi de probabilité fixée à l'avance à support dans J_N.

On a, pour tout p, $T_N(p)=1/N\,\Sigma(pA_i/(pA_i+(1-p)B_i);i=1,N)$

Considérons, d'autre part, la fonction définie sur $[0,1]$:

$T(p)=\int(pf_1(x)/pf_1(x)+(1-p)f_2(x))f^*(x)dx$

$f^*(x)$ étant la vraie densité du mélange.

Notons p_S^N une v.a. de loi la loi stationnaire de la chaine de Markov homogène ergodique associé à l'algorithme SEM.

Pour démontrer la convergence de la loi stationnaire de la suite $(p_S^N-p^*)\sqrt{N}$ vers une loi normale centrée nous allons utiliser les quatre lemmes suivants.

Lemme 1: (cf <Si80>)

$T(0)=0$, $T(1)=1$, $T(p^*)=p^*$
Pour $0<p<p^*$ $T(p)>p$
Pour $p^*<p<1$ $T(p)<p$

Lemme 2:

$T'(0)>1$, $T'(1)>1$, $T'(p^*)<1$

Démonstration:

On déduit immédiatement du lemme 1 que $T'(0)\geqslant 1$, $T'(1)\geqslant 1$ et $T'(p^*)\leqslant 1$. En dérivant sous le signe $\int$ il vient:

$$T'(0)=\int f^*(x)(f_1(x)/f_2(x))dx, T'(1)=\int f^*(x)(f_2(x)/f_1(x))dx$$

Montrons que $T'(0)$ est strictement plus grand que 1.

$$T'(0)=p^*\int f_1^2(x)/f_2(x)dx+(1-p^*)$$

Par l'inégalité de Jensen, on a:

$$\int f_1^2(x)/f_2(x)dx > 1/\int (f_2(x)/f_1(x))f_1(x)dx = 1$$

Cette inégalité étant stricte car $f_1\neq f_2$. On a donc:

$$T'(0)>p^*+(1-p^*)=1$$

De manière analogue, on montre que $T'(1)>1$.

Enfin la démonstration du fait que $T'(p^*)$ est strictement plus petite que 1 est exactement analogue à la démonstration, détaillée dans la proposition 1 ci-après, du fait que $T'_N(p_N)$ est strictement plus petit que 1. Nous invitons le lecteur à s'y reporter. C.Q.F.D.

Lemme 3:

1) Pour tout p appartenant à $[0,1]$ $T_N(p)$ converge p.s. vers $T(p)$.

2) Pour tout p appartenant à $]0,1[$ $T'_N(p)$ converge p.s. vers $T'(p)$

et $T''_N(p)$ converge p.s. vers $T''(p)$.

Démonstration:

Il suffit de vérifier que les hypothèses de la loi forte des

grands nombres (théorème de Kolmogorov, voir par exemple <Fe71> Théorème 1 page 238) sont satisfaites.

1) $T_N(p)=1/N\,\Sigma(pf_1(x_i)/(pf_1(x_i)+(1-p)f_2(x_i));i=1,N)$

Les v.a. sous le signe Σ sont équidistribuées et indépendantes puisque ce sont des fonctions mesurables de v.a. équidistribuées et indépendantes.
Elles ont pour espérance T(p) et:

$$T(p)\leqslant\int f^*(x)dx=1$$

2) $T'_N(p)=1/N\,\Sigma(f_1(x_i)f_2(x_i)/(pf_1(x_i)+(1-p)f_1(x_i))^2;i=1,N)$

$T''_N(p)=1/N\,\Sigma(-2f_1(x_i)f_2(x_i)(f_1(x_i)-f_2(x_i))/(pf_1(x_i)+(1-p)f_1(x_i))^3;i=1,N)$

De la même facon qu'en 1) les v.a. sous le signe Σ, dans ces deux expressions, sont équidistribuées et indépendantes et ont pour espérances respectives T'(p) et T''(p). Par ailleurs, pour $p\neq 0$ et $p\neq 1$, on a:

$$T'(p)=\int f^*(x)f_1(x)f_2(x)/(pf_1(x)+(1-p)f_2(x))^2dx$$

$$T'(p)=(1/(p(1-p)))\int f^*(x)pf_1(x)(1-p)f_2(x)/(pf_1(x)+(1-p)f_2(x))^2dx$$

d'où $0<T'(p)\leqslant 1/(p(1-p))$

Et:

$$T''(p)=-2\int f^*(x)f_1(x)f_2(x)(f_1(x)-f_2(x))/(pf_1(x)+(1-p)f_2(x))^3dx$$

$|T''(p)|=2/(p^2(1-p)^2)I$ avec:

$$I=\int f^*(x)pf_1(x)(1-p)f_2(x)p(1-p)|f_1(x)-f_2(x)|/(pf_1(x)+(1-p)f_2(x))^3dx$$

Soit $A=\{x/f_1(x)>f_2(x)\}$ et $B=\{x/f_1(x)\leqslant f_2(x)\}$

$$|T''(p)|<2/(p^2(1-p))\int_A f^*(x)dx+2/((1-p)^2p)\int_B f^*(x)dx$$

$|T''(p)|<2/(p^2(1-p))+2/((1-p)^2p)$ C.Q.F.D.

<u>Lemme 4</u>:
Il existe N_0 tel que pour tout $N>N_0$ $T'_N(0)>1$ et $T'_N(1)>1$ p.s.

Démonstration:
Si T'(0) et T'(1) sont finis, le 2) du lemme 3 s'étend aux cas p=0 et p=1. Le résultat annoncé vient alors du lemme 2.

Si $T'(0)$ (resp.$T'(1)$)$=\infty$, $T'_N(0)$ (resp.$T'_N(1)$) est non bornée p.s. (cf <Fe71> Théorème 4 page 241) C.Q.F.D.

Proposition 1:

1) Pour N suffisamment grand, T_N admet un unique point fixe p_N compris strictement entre 0 et 1. De plus,

$0<T'_N(p_N)<1$.

2) Ce point fixe réalise un maximum global unique de la vraisemblance de l'échantillon de taille N sur $[0,1]$.

3) La suite des p_N converge p.s. vers p^*.

Démonstration:

1) D'après le lemme 4, pour N suffisamment grand, on a $T'_N(0)>1$ et $T'_N(1)>1$. On se place désormais dans ce cas.

Des formules:

$T'_N(p)=1/N\,\Sigma(A_iB_i/(pA_i+(1-p)B_i)^2);i=1,N)$

$T'''_N(p)=1/N\,\Sigma(-2A_iB_i(A_i-B_i)^2/(pA_i+(1-p)B_i)^4);i=1,N)$,

on déduit que T_N et T''_N sont strictement croissantes. Il s'en suit que:

$T''_N(0)<0$ et $T''_N(1)>0$.

En effet:

Si $T''_N(0)\geqslant 0$, la fonction T_N serait convexe sur $[0,1]$ car T''_N est strictement croissante.

Donc le graphe de T_N serait au dessus de la tangente au point (0,0) et ne contiendrait pas le point (1,1) puisque cette tangente a une pente $T'_N(0)$ stritement plus grande que 1.

De manière analogue, on montre que $T''_N(1)$ est strictement positive.

Par ailleurs, T''_N s'annule en un point unique p_i sur $]0,1[$ puisque T''_N est continue et strictement croissante.

Montrons maintenant l'existence et l'unicité du point fixe p_N.

Existence:

De $T'_N(0)>1$, on déduit qu'il existe des points p de $]0,1[$ tels que

$T_N(p)/p>1$, i.e. tels que $T_N(p)>p$.

De $T'_N(1)>1$, on déduit qu'il existe des points p de $]0,1[$ tels que

$(1-T_N(p))/(1-p)>1$, i.e. tels que $T_N(p)<p$.

La fonction T_N étant continue, il s'en suit qu'elle possède au moins un point fixe sur $]0,1[$.

Unicité:

Soit p_1 le plus petit point fixe strictement positif. Deux cas sont à étudier.

Premier cas: $p_i \leqslant p_1$.

Alors, le graphe de T_N est convexe sur $[p_1,1]$. Donc la première bissectrice rencontre ce graphe en deux points au plus qui sont (p_1,p_1) et $(1,1)$.

Deuxième cas: $p_i>p_1$.

D'après le théorème des accroissements finis, il existe c strictement compris entre 0 et p_1 tel que $T'_N(c)=1$. T'_N est strictement décroissante sur $[0,p_i[$ et donc

$0<T'_N(p_1)<1$.

T_N étant concave sur $[0,p_i]$ son graphe dans cet intervalle est situé sous la tangente au point (p_1,p_1) et donc pour tout p appartenant à $]p_1,p_i]$ on a $T_N(p)<p$.

Soit p_2 le deuxième point fixe de T_N $(p_2>p_i)$. D'après le théorème des accroissements finis, il existe d strictement compris entre p_i et p_2 tel que $T'_N(d)>1$ car $T_N(p_i)<p_i$.

Si $p_2<1$, le graphe de T_N ne contiendrait pas le point $(1,1)$ et donc $p_2=1$.

En effet $(1,1)$ est au dessous de la tangente au point (p_2,p_2) et T_N étant convexe sur $[p_i,1]$ son graphe, dans cet intervalle, est situé au dessus de cette tangente.

Montrons maintenant que $T'_N(p_N)<1$.

On vient de le voir dans le cas où $p_i>p_N$.

Si $p_i \leqslant p_N$, supposons que $T'_N(p_N) \geqslant 1$.

Le graphe de T_N, convexe sur $[p_N,1]$, serait situé au dessus de la tangente au point (p_N,p_N) et donc ne contiendrait pas les points du graphe en dessous de la première bissectrice (qui existent du fait que $T'_N(1)>1$).

2) La Log-vraisemblance s'écrit pour l'échantillon de taille N:

$L_N(p)=\Sigma(\mathrm{Log}(pA_i+(1-p)B_i);i=1,N)$

$L'_N(p)=\Sigma((A_i-B_i)/(pA_i+(1-p)B_i);i=1,N)$

On a:

$L'_N(0)=\Sigma(A_i/B_i;i=1,N)-N=N(T'(0)-1)>0$ (lemme 4)

$L'_N(1)=N-\Sigma(B_i/A_i;i=1,N)=N(1-T'(1))<0$ (lemme 4)

Si maintenant $p\neq 0$ et $p\neq 1$, on a:

$p(1-p)L'_N(p)=N(T_N(p)-p)$

p_N est donc le seul point de $[0,1]$ vérifiant $L'_N(p_N)=0$.

p_N réalise bien le maximum global unique de la vraisemblance sur $[0,1]$ car:

$L''_N(p)=-\Sigma((A_i-B_i)^2/(pA_i-(1-p)B_i)^2;i=1,N) < 0$ pour tout p.

3) Du fait que p_N réalise le maximum global unique de la vraisemblance, on déduit que p_N converge p.s. vers p^* (cf <KeSt73> paragraphe 18.10 pages 41 à 43). C.Q.F.D.

Dans la suite de ce paragraphe, nous aurons besoin du résultat suivant.

<u>Proposition 2:</u>

$|T'_N(p_N)$ converge p.s. vers $T'(p^*)$ noté ρ^*.

Démonstration:

On a l'inégalité (1):

$|T'_N(p_N)-T'(p^*)|<|T'_N(p_N)-T'_N(p^*)|+|T'_N(p^*)-T'(p^*)|$

D'une part, $T'_N(p^*)$ converge p.s. vers $T'(p^*)$ d'après le lemme 3 puisque $0<p^*<1$.

D'autre part, comme p_N converge p.s. vers p^*, il existe p.s. un intervalle $I=[u,v]$ contenant p^* avec $u>0$, $v<1$ et indépendant de N tel que, pour N suffisamment grand, $p_N \in I$. On se place évidemment dans ce cas.

On a vu au cours de la démonstration de la proposition 1 que T''_N était strictement croissante. D'où:

Pour tout $p \in I$, $|T''_N(p)| < \sup(|T''_N(u)|, |T''_N(v)|)$

Et on montre de manière analogue à la majoration de $T''(p)$ dans le lemme 3 que

$|T''_N(u)| < 2/(u^2(1-u)) + 2/(u(1-u)^2)$

$|T''_N(v)| < 2/(v^2(1-v)) + 2/(v(1-v)^2)$

Donc $M=\sup(|T''_N(p)|\,;p\in I)$ est un nombre fini indépendant de N.

D'après le théorème des accroissements finis, on a:

$|T'_N(p_N)-T'_N(p^*)| < M|p_N-p^*|$

D'après la proposition 1, p_N converge p.s. vers p^*. Donc l'expression

$|T'_N(p_N)-T'_N(p^*)|$

peut être majorée p.s., pour N assez grand, par un nombre arbitrairement petit.

On en déduit immédiatement, d'après l'inégalité (1), le résultat annoncé. C.Q.F.D.

4.2. Modélisation approximative de l'algorithme SEM

Soit $c(N)$ une suite de réels de $]0,1[$ tendant vers 0 lorsque N tend vers l'infini, ou bien restant constante.

Soit J_N l'intervalle $[c(N),1-c(N)]$.

Définissons par ailleurs la fonction s_N par:

$s_N^2(p)=1/N\,\Sigma(t(x_i;p)(1-t(x_i;p));i=1,N)$ si $p \in J_N$

$s_N^2(p)=$Cste positive ou nulle sinon.

$t(x_i;p)$ représentant la probabilité a posteriori, associée à p, d'appartenance de x_i au premier composant.

On a sur J_N, $s_N^2(p)=p(1-p)T'_N(p)$.

Notons $\rho_N=T'_N(p_N)$. Soit enfin:

$\eta^N(e;p)=(\Sigma\,(t(x_i;p)(1-t(x_i;p));i=1,N))^{-1/2}S_N(p)$ avec:

$S_N(p)=\Sigma(e(x_i;p)-t(x_i;p);i=1,N)$

De sorte que $E_e(\eta^N)=0$ et $E_e((\eta^N)^2)=1$.

Nous avons décrit au début du paragraphe 4.1. les équations qui décrivent le fonctionnement de l'algorithme SEM dans le cas particulier étudié ici.

L'étude du comportement asymptotique lorsque N tend vers l'infini de la loi stationnaire ψ^N de la chaine associée est mal commode: il est plus simple de remplacer ces équations, dans un premier temps, par la seule équation suivante qui décrit une chaine de Markov homogène $(X_n^N)_n$, dont nous montrons l'ergodicité au paragraphe 4.3:

(M) $X_{n+1}^N=T_N(X_n^N)+N^{-1/2}s_N(X_n^N)\eta_{n+1}^N(e;X_n^N)$

où T_N et s_N ont été modifiés de manière à coincider sur J_N avec les fonctions T_N et s_N initiales et à être constantes à l'extérieur de J_N,

la première constante $T_N(c-)=T_N((1-c)+)$

appartenant à $]c(N),1-c(N)[$,

la seconde $s_N(c-)=s_N((1-c)+)$ étant positive ou nulle.

Ainsi, si $X_{n+1}^N \notin J_N$, alors

$T_N(X_{n+1}^N)=T_N(c-) \in\ ]c(N),1-c(N)[$, et $s_N(X_{n+1}^N)=S_N(c-)$.

De sorte que si $s_N(c-)$ est stritement positif X_{n+2}^N est tirée selon

la loi normale de moyenne $T_N(c-)$ et d'écart-type $s_N(c-)$

ou selon la mesure de Dirac au point $T_N(c-)$ si $s_N(c-)$ est nul.

Nous nous proposons:

- de montrer l'ergodicité de la chaine de Markov (M) dont on notera φ^N la loi stationnaire.

-de démontrer le résultat asymptotique suivant:

Si l'on note X^N une v.a. de loi la loi stationnaire φ^N

et $Y^N=N^{1/2}(X^N-p_N)$,

alors quand N tend vers l'infini, la loi de Y^N converge vers la loi normale centrée d'écart-type σ avec:

$\sigma^2=s^2(p^*)(1-(\rho^*)^2)^{-1}$

où $\rho^*=T'(p^*)<1$ (cf lemme 2).

-de montrer que si l'on choisit, pour l'algorithme SEM, une loi de relance induite par le modèle (M) dans le cas où le tirage aléatoire fournit un résultat

X_n^N à l'extérieur de J_N,

si l'on note p_S^N une v.a. dont la loi est la loi stationnaire ψ^N de l'algorithme ainsi relancé, si l'on note

$q_S^N=N^{1/2}(p_S^N-p_N)$,

alors la loi de q_S^N est "équivalente" en un sens précisé au théorème 2

quand N tend vers l'infini, à celle de Y^N.

Il en résultera alors que, pour N assez grand, on peut considérer que la loi stationnaire de la suite $((p_S^N)_n,n>0)$ engendrée par l'algorithme SEM est approximativement une loi normale de moyenne p_N et d'écart-type $N^{-1/2}\sigma$.

Pour simplifier les calculs, on sera amené à considérer des notations centrées. On notera:

$\bar{X}^N_n=X_n^N-p_N$

$\bar{T}_N(q)=T_N(q+p_N)-p_N$

$\bar{s}_N(q)=s_N(q+p_N)$

$J_N=[g(N),h(N)]$ avec $g(N)=c(N)-p_N<0$ et $h(N)=c(N)+p_N>0$

Avec ces notations, le modèle (M) s'écrit:

$\bar{X}^N_{n+1}=\bar{T}_N(\bar{X}^N_n)+N^{-1/2}\bar{s}_N(\bar{X}^N_n)\eta^N_{n+1}(e;\bar{X}^N_n)$

Par ailleurs, on déduit facilement de l'étude de la fonction T_N (cf paragraphe 4.1) que pour N assez grand il existe un nombre R(N) strictement compris entre 0 et 1 tel que:

$p \in J_N \Longrightarrow |T_N(p)-p_N| < R(N)|p-p_N|$

En notations centrées , cette inégalité devient:

$|\bar{T}_N(q)| < R(N)|q|$.

4.3. Ergodicité de la chaine (M)

Proposition 3:

Pour tout N fixé, la chaine (M) est ergodique.

Démonstration:

Montrons que (M) est irréductible pour la mesure de comptage ν (cf <Re75> page 71).

-Si $X_n^N=p \in J_N$, $X_{n+1}^N=N^{-1}\Sigma(e(x_i);i=1,N)$ ne peut prendre que les valeurs

$0,1/N,\ldots,(N-1)/N,1$;

notons F cet ensemble.

-Si $X_n^N=p>1-c(N)$

$X_{n+1}^N=T_N((1-c)+)+N^{-1}\Sigma(e(x_i;1-c(N))-t(x_i;1-c(N));i=1,N)$

X_{n+1}^N appartient encore à un ensemble fini G;

de plus la probabilité que tous les X_{n+m}^N $(m>1)$ appartiennent à $G\cap]1-c(N),+\infty[$ est nulle.

-De même si $X_n^N=p<c(N)$, X_{n+1}^N appartient à un ensemble fini H.

c(N) étant strictement positif, la probabilité que les X_{n+m}^N $(m>1)$ restent dans J_N sachant que

$X_n^N \in J_N$

est également nulle.

Ainsi, si E désigne l'ensemble fini

$(F\cap J_N)\cup(G\cap]1-c(N),+\infty[)\cup(H\cap]-\infty,c(N)[)$ et

ν la mesure de comptage normalisée de E, si $A\in E$ vérifie

$\nu(A)>0$ (i.e. $A\neq\Phi$),

si $X_n^N=p$, la probabilité pour qu'il existe $m>1$ tel que $X_{n+m}^N \in A$ est non nulle. (M) est donc ν-irréductible.

Ainsi la chaine se ramène à une chaine à nombre fini d'états et irréductible, donc elle est ergodique.

4.4. Convergence en loi du bruit η^N vers un bruit blanc Gaussien

4.4.1. Théorème central limite uniforme en $p \in J_N$

Précisons d'abord l'espace probabilisé des échantillons: c'est le produit tensoriel d'une infinité dénombrable, indexée par $n>1$, d'exemplaires de $(\mathbb{R},\mathcal{B}(\mathbb{R}),f^*(x)dx)$. Chaque élément est une suite $\underline{x}=(x_n;n>1)$.

Précisons aussi que, dans le modèle (M), les entrées sont les v.a. $\eta^N(e;p)$ avec

$\eta^N(e;p)=\eta^N(e;1-c(N))$ pour $p>1-c(N)$

et $\eta^N(e;p)=\eta^N(e;c(N))$ pour $p<c(N)$.

On notera:

$e_m(x_i;p)=e(x_i;p)-t(x_i;p)$

$S_N(e;p)=\Sigma(e_m(x_i;p);i=1,N)$

et on a:

$s_N(p)=E.T.(S_N(e;p))$, $\eta^N=s_N^{-1}S_N$.

On note E.T.(X)=écart-type(X)

Proposition 4:

Supposons que la suite $(c(N);N>0)$ est constante ou tend vers 0, lorsque N tend vers l'infini, de telle sorte que $Nc(N)$ tend vers l'infini.

Alors, pour tout échantillon $\underline{x}$ n'appartenant pas à un ensemble de mesure nulle, pour tout $a \in R$, pour tout $\varepsilon>0$, il existe un entier $N_0(\underline{x},a,\varepsilon)$ tel que:

$N>N_0 \Longrightarrow (\forall p)\ |P(\eta^N(e;p)<a)-\Phi(a)|<\varepsilon,$

où Φ désigne la f.d.r. de la loi normale centrée réduite.

Démonstration:

Dans l'étape 1, nous montrerons que, $\underline{x}$ p.s., $s_N(p)$ tend vers l'infini assez vite avec N; dans l'étape 2, nous montrons, comme dans <Bi68> p42-44, que la suite

$s_N^{-1}S_N(e;p)$

converge étroitement, uniformément en p, lorsque N tend vers l'infini; dans l'étape 3, nous en déduisons, en reprenant la preuve du théorème du portemanteau de <Bi68> p12-14, le résultat de la proposition.

Etape 1: Par la loi des grands nombres, on a, $\underline{x}$ p.s.,

$\lim N^{-1}s_N^2(p)=p(1-p)\int f_1(t)f_2(t)f^*(t)/(pf_1(t)+(1-p)f_2(t))^2dt \quad p \in J_N.$

Si $c(N)\leqslant 1/2$ comme $p\geqslant c(N)$, on a

$\lim N^{-1}s_N^2(p)\geqslant c(N)(1-c(N))\int f_1(t)f_2(t)f^*(t)/(f_1(t)+f_2(t))^2dt$

L'intégrale ci-dessus converge car $f_1f_2/(f_1+f_2)^2\leqslant 1$. Soit I sa valeur.

On a donc $s_N^2(p)>c(N)NI(1-c(N))$

On en déduit que $s_N^2(p)$ tend vers l'infini car par hypothèse $c(N)N$ tend vers l'infini avec N.

Etape 2: Comme dans <Bi68>, soit f une fonction C^∞, dont toutes les dérivées sont bornées, soit $K=K(f)$ une constante telle que si g désigne le reste de la formule de Taylor-Lagrange d'ordre 2 appliquée à f, on ait:

pour tout h, $|g(h)|<K\min(|h|^2,|h|^3)$.

Introduisons des v.a. normales centrées réduites indépendantes entres elles et indépendantes des $e_m(x_i;p)$, que nous noterons G_{Ni}, $i=1,N$;

notons $s_{Ni}(p)=E.T.(e_m(x_i;p))$, de sorte que l'on a:

$s_N^2(p)=\Sigma(s_{Ni}^2(p);i=1,N)$

On montre dans <Bi68> que, si G est une v.a. normale centrée réduite

$|Ef(s_N(p)^{-1}S_N(e;p))-Ef(G)|<\Sigma(E|g(s_N(p)^{-1}e_m(x_i;p))|;i=1,N)$

$+\Sigma(E|g(s_N(p)^{-1}s_{Ni}(p)G_{Ni})|;i=1,N)$

a) La première somme est majorable par

$$K\eta+s_N(p)^{-2}K\Sigma\left(\int_{\{|e_m(x_i;p)|>\eta s_N(p)\}}|e_m(x_i;p)|^2dP;i=1,N\right) \text{ pour } \eta>0 \text{ donné;}$$

Or, les v.a. $|e_m(x_i;p)|$ sont plus petites ou égales à un; puisque $s_N(p)$ tend vers l'infini quand N tend vers l'infini uniformément en p, les ensembles

$$\{|e_m(x_i;p)|>\eta s_N(p)\}$$

sont tous vides pour N assez grand: il reste une majoration par $K\eta$.

b) De même, la seconde somme est majorée par:

$$K\eta+K\Sigma\left(s_N(p)^{-2}s_{Ni}(p)^2\int_{\{|G_{Ni}|>\eta s_N(p)\}}|G_{Ni}|^2dP;i=1,N\right)$$

Soit G une v.a. normale centrée réduite, on a:

$$\int_{\{|G|>\eta s_N\}}|G|^2dP=\int_{\{1<\eta^{-1}s_N(p)^{-1}|G|\}}|G|^2dP$$

$$\leqslant\int_{\{1<\eta^{-2r}s_N(p)^{-2r}|G|^{2r}\}}|G|^2dP$$

$$\leqslant\int\eta^{-2r}s_N(p)^{-2r}|G|^{2r+2}dP$$

$$<Cste(r)\eta^{-2r}(c(N)NI(1-c(N)))^{-r} \qquad (r \text{ fixé})$$

Cette quantité tend vers 0 quand N tend vers l'infini car c(N)N tend vers l'infini.

Etape 3: Nous déduisons de la preuve du théorème 2.1 de <Bi68> que, si une suite de v.a. réelles converge étroitement, uniformément en p, vers une v.a. normale centrée réduite, alors, puisque tout a R est point de continuité de Φ, la suite des f.d.r. converge, uniformément en p, vers Φ. C.Q.F.D.

4.4.2. Mise en forme d'un principe d'invariance

Ce principe est dû à Skorohod (<Sk56>): il s'agit, sachant que η^N converge en loi vers un bruit blanc gaussien uniformément en $p\in J_N$,

de construire sur un espace probabilisé adéquat, des suites indépendantes et équidistribuées de v.a.

$(\hat{\eta}_n^N(\underline{w};p);n>0)$ et $(\hat{\varepsilon}_n(\underline{w});n>0)$ telles que:

(C) $(\forall n>0)$ $(\forall p\in J_N)$ la suite (en N) $\hat{\eta}_n^N(\underline{w};p)$ converge p.s. vers $\hat{\varepsilon}_n(\underline{w})$, la convergence étant uniforme en p.

On montrera alors que:

-Si X_o^N est une v.a. de loi la loi stationnaire φ^N de la chaine (M);

-si $Y_o^N=N^{1/2}\bar{X}_o^N$;

-si la suite $(Z_n^N;n>0)$ est définie par:

(MG) $Z_{n+1}^N=\rho_N Z_n^N+\bar{s}_N(0)\hat{\varepsilon}_{n+1}(\underline{w})$ et $Z_o^N=Y_o^N$, alors:

(i) Uniformément en n, la limite lorsque N tend vers l'infini de:

$E^{1/2}|Y_n^N-Z_n^N|^2$ est nulle.

(ii) Si la suite $(Z^*_n;n>0)$ est définie par:

(MG*) $Z^*_{n+1}=\rho^*Z^*_n+s^*\hat{\varepsilon}_{n+1}(\underline{w})$ et $Z^*_o=0$

s^* étant la limite de $s_N(p_N)$ lorsque N tend vers l'infini;

alors il existe une suite V(N) tendant vers 0, telle que:

$E^{1/2}|Y_n^N-Z^*_n|^2\leqslant(1-R^2(N))^{-1/2}\rho^{*n}+V(N)$

(iii) Notons $\Phi_{o,\sigma}$ la f.d.r. de la loi normale centrée d'écart-type σ, avec:

$\sigma^2=s^{*2}(1-\rho^{*2})^{-1}$

Par choix d'une suite n(N) tendant vers l'infini, il en résultera que:

$\lim\ (P(\bar{X}^N N^{1/2}<a)-\Phi_{o,\sigma}(a))=0$

car $\Phi_{o,\sigma}(a)=\lim\ P(Z^*_{n(N)}<a)$.

Procédons à la construction de $\hat{\eta}$ et $\hat{\varepsilon}$, en suivant <Br68> p 293-297:

Si F désigne une fonction de répartition, notons

$F^{-1}(x)=\inf(y/F(y)>x)$

Notons λ la mesure de Lebesgue normalisée sur $[0,1]$.

Formons $W=[0,1]^{N*}$, $\mathcal{W}=\otimes_{N*}\mathcal{B}[0,1]$, $\Lambda=\otimes_{N*}\lambda$;

On notera $\underline{w}$ un élément de W: $\underline{w}=(w_1,\ldots,w_n,\ldots)$

Notons $F_N(.,p)$ la fonction de répartition de la v.a. $\eta^N(e;p)$ $(p\in J_N)$ et $F_N^{-1}(.,p)$ la fonction définie ci-dessus.

Notons Φ la f.d.r de la loi normale centrée réduite.

Formons $\hat{\eta}_n^N(\underline{w};p)=F_N^{-1}(w_n;p)$ et $\hat{\varepsilon}_n(\underline{w})=\Phi^{-1}(w_n)$

Ces suites de v.a. indépendantes et équidistribuées vérifient bien (C).
On omettra par la suite ces notations particulières $\hat{\eta}$ et $\hat{\varepsilon}$ pour revenir à η et ε.

4.5. Théorème principal de convergence

Théorème 1:

Si c(N) tend vers 0 assez lentement quand N tend vers l'infini

si X^N est une v.a de loi la loi stationnaire de la chaine (M), on a pour tout a:

$$\lim P(N^{1/2}(X^N-p_N)<a)=\Phi_{0,\sigma}(a)$$

où $\Phi_{0,\sigma}$ désigne la f.d.r. de la loi normale centrée d'écart-type σ, avec

$$\sigma^2=s^{*2}(1-\rho^{*2})^{-1}.$$

Lemme 5

Soit X^N une v.a. dont la loi est la loi stationnaire de la chaine (M) centrée,

soit $Y^N=N^{1/2}\bar{X}^N$

(i) $E(Y^N)^2<(1-R(N)^2)^{-1}$

(ii) $E(Y^N)^4<Cste(1-R(N))^{-4}$

Démonstration:

(i) Conditionnons par $\bar{X}^N_n=q$, et supposons que $loi(\bar{X}^N_n)=\varphi^N$:

$$E(|\bar{X}^N_{n+1}|^2/\bar{X}^N_n=q)=\bar{T}_N^2(q)+N^{-1}\bar{s}_N^2(q)$$

(car $E_e\eta=0, E_e\eta^2=1$)

Utilisant $\bar{T}_N^2(q)<R(N)^2q^2$ et déconditionnant:

$E|\bar{X}^N_{n+1}|^2\leqslant R(N)^2\int q^2\varphi^N(dq)+N^{-1}$ (car $\bar{s}_N^2(q)<1$)

Par stationnarité: $E(|\bar{X}^N|^2)\leqslant N^{-1}(1-R(N)^2)^{-1}$.

(ii) Par l'inégalité de Minkowski, il vient:

$E^{1/4}|\bar{X}^N_{n+1}|^4<R(N)E^{1/4}|\bar{X}^N_n|^4+N^{-1/2}E^{1/4}|\eta^N_{n+1}(e;\bar{X}^N_n)|^4$

$E^{1/4}|\bar{X}^N_{n+1}|^4<R(N)E^{1/4}|\bar{X}^N_n|^4+CsteN^{-1/2}$

(car d'après la proposition 4, $E|\eta^N(e;q)|^4$ tend vers $E|\epsilon|^4$ uniformément en $q\in\bar{J}_N$)

Par stationnarité:

$E|\bar{X}^N|^4<CsteN^{-2}(1-R(N))^{-4}$. C.Q.F.D.

<u>Lemme 6</u>

Notons $B(N)=\sup(1,\max(|T''_N(c)|,|T''_N(1-c)|)$

et $G(N)=\sup(1,\sup_{J_N}|s_N'|)$

(i) il existe C_1 et N_1 tels que $N>N_1$ entraine:

$|N^{1/2}\bar{T}_N(qN^{-1/2})-\rho_Nq|<C_1q^2N^{-1/2}B(N)$

(ii) il existe C_2 et N_2 tels que $N>N_2$ entraine:

$|\bar{s}_N(qN^{-1/2})-\bar{s}_N(0)|<C_2|q|N^{-1/2}G(N)$

Démonstration:

(i) Cas où $qN^{-1/2}\in\bar{J}_N$: On a

$N^{1/2}\bar{T}_N(qN^{-1/2})=\rho_Nq+(q^2/2)N^{-1/2}\bar{T}''_N(\theta(qN^{-1/2}))$ avec $\theta\in[0,1]$.

d'après la formule de Taylor-Lagrange à l'ordre 2.

D'où $|N^{1/2}\bar{T}_N(qN^{-1/2})-\rho_Nq|<q^2N^{-1/2}\sup_{\bar{J}_N}|\bar{T}''_N|$

Cas où $qN^{-1/2}\notin\bar{J}_N$; supposons $qN^{-1/2}>h(N)=\sup\bar{J}_N$:

On a alors , par définition de T_N,

$|N^{1/2}\bar{T}_N(qN^{-1/2})|=N^{1/2}|\bar{T}_N(h+)|<N^{1/2}$;

et, d'autre part $CN^{-1/2}q^2 > CN^{-1/2}h(N)^2 N > N^{1/2}$

dès que $Ch(N)^2 \geqslant 1$;

Enfin toujours dans ce cas, on a, de même,

$CN^{-1/2}q^2 \geqslant \rho_N q^2 N^{-1/2} h(N)^{-1}$ si $\rho_N h(N)^{-1} \leqslant C$

$\geqslant (\rho_N q N^{-1/2} h(N)^{-1}) N^{1/2} h(N) = \rho_N q$ car $q \geqslant h(N) N^{1/2}$

Or, quand N tend vers l'infini, h(N) tend vers 1-p* si c(N) tend vers 0 et vers 1-c(N)-p* si c(N) reste constant.
Donc, dans ce cas, il existe une constante C_1 et un entier N_1 tels que $N \geqslant N_1$ entraine:

$|N^{1/2}\bar{T}_N(qN^{-1/2}) - \rho_N q| < C_1 N^{-1/2} q^2$

On traite de manière analogue le cas où $qN^{-1/2} < g(N) = \inf \bar{J}_N$.

(ii) se traite de manière analogue. C.Q.F.D.

Démonstration du théorème 1:

Etape 1:

On va montrer que si X_n^N vérifie (M) et est stationnaire, si

$Y_n^N = \bar{X}^N{}_n N^{1/2}$ pour $n>0$, $Y_o^N = Z_o^N$

et $Z_{n+1}^N = \rho_N Z_n^N + \bar{s}_N(0)\varepsilon_{n+1}(e)$

alors on a (pour tout $n>0$): $E|Y_n^N - Z_n^N|^2 < U^2(N)$

avec lim U(N)=0. Posons:

$A(Y_n^N) = N^{1/2}\bar{T}_N(Y_n^N N^{-1/2}) - \rho_N Y_n^N$, $D(Y_n^N) = \bar{s}_N(Y_n^N N^{-1/2}) - \bar{s}_N(0)$

On a:

$Y_{n+1}^N = \rho_N Y_n^N + A(Y_n^N) + \bar{s}_N(0)\ \eta_{n+1}^N(e, Y_n^N N^{-1/2}) + D(Y_n^N)\ \eta_{n+1}^N(e, Y_n^N N^{-1/2})$

Par le lemme 6, on a:

$|A(Y_n^N)| < C_1 N^{-1/2} (Y_n^N)^2 B(N)$

$|D(Y_n^N)| < C_2 N^{-1/2} |Y_n^N| G(N)$

On choisit conformément à la construction du paragraphe 4.4.2 une suite

$(\eta_n^N(\underline{w}; q); n>0)$

qui converge p.s. uniformément en $q \in J_N$, vers la suite

$(\varepsilon_n(\underline{w});n>0)$.

Par l'inégalité de Minkowski, on a:

$$E^{1/2}|Y^N_{n+1}-Z^N_{n+1}|^2 \leqslant \rho_N E^{1/2}|Y^N_n-Z^N_n|^2+B(N)C_1N^{-1/2}E^{1/2}|Y^N_n|^4$$

$$+\bar{s}_N(0)E^{1/2}|\eta^N_{n+1}(\underline{w};Y^N_nN^{-1/2})-\varepsilon_{n+1}(\underline{w})|^2$$

$$+G(N)C_2N^{-1/2}E^{1/2}|Y^N_n|^2\ E^{1/2}|\eta^N_{n+1}(\cdot;q)|^2\varphi^N(dq).$$

Par le lemme 5, ceci est inférieur ou égal à:

$$E^{1/2}|Y^N_n-Z^N_n|^2+B(N)C_1N^{-1/2}Cste(1-R(N))^{-2}$$

$$+\bar{s}_N(0)E^{1/2}|\eta^N_{n+1}(\underline{w};Y^N_nN^{-1/2})-\varepsilon_{n+1}(\underline{w})|^2$$

$$+G(N)C_2N^{-1/2}(1-R^2(N))^{-1/2}\int\varphi^N(dq).$$

Notons $u^N_n=E^{1/2}|Y^N_n-Z^N_n|^2$ et $r(N)$ le reste dans l'expression ci-dessus, on obtient:

$$u^N_{n+1}<\rho_N u^N_n+r(N) \quad \text{et} \quad u^N_o=0,$$

avec $\lim r(N)=0$ (d'après la proposition 4) pourvu que la suite $c(N)$ tende vers 0 quand N tend vers l'infini assez lentement pour que:

$$\lim N^{-1/2}B(N)(1-R(N))^{-1}=0$$

$$\lim N^{-1/2}G(N)(1-R^2(N))^{-1/2}=0$$

Posons $u^N=\rho_N u^N+r(N)$ d'où $u^N=r(N)(1-\rho_N)^{-1}$

avec $\lim \rho_N=\rho^*$ (proposition 2)

On montre par récurrence que $u^N_n<u^N$.

Etape 2:

Comparons Z^N_n engendrée par (MG) et Z^*_n engendrée par (MG*). On a:

$$Z^N_{n+1}-Z^*_{n+1}=\rho_N Z^N_n-\rho^*Z^*_n+(\bar{s}_N(0)-s^*)\varepsilon_{n+1} \text{ donc:}$$

$$E^{1/2}|Z^N_{n+1}-Z^*_{n+1}|^2 \leqslant |\rho_N-\rho^*|E^{1/2}|Z^N_n|^2+\rho^*E^{1/2}|Z^N_n-Z^*_n|^2+|\bar{s}_N(0)-s^*|$$

$$\leqslant \rho^*E^{1/2}|Z^N_n-Z^*_n|^2+r'(N)$$

avec lim $r'(N)=0$ et avec

$$E^{1/2}|Z_o^N-Z^*_o|^2=E^{1/2}|Y_o^N|^2<(1-R^2(N))^{-1/2}$$

d'après le lemme 5.

Notons $v_n^N=E^{1/2}|Z_n^N-Z^*_n|^2$, on obtient:

$$v_n^N<\rho^{*n}(1-R^2(N))^{-1/2}+(1-\rho^*)^{-1}r'(N)$$

Etape 3:

Il en résulte que:

$$E^{1/2}|Y_n^N-Z^*_n|^2<\rho^{*n}(1-R^2(N))^{-1/2}+V(N)$$

avec lim $V(N)=0$.

Il est possible de choisir la suite $n(N)$ tendant vers l'infini pour que:

$$\lim \rho^{*n(N)}(1-R^2(N))^{-1/2}=0$$

En effet, le log de cette expression s'écrit:

$$n(N)\log\rho^*+1/2|\log(1-R^2(N))|$$

qui tend vers moins l'infini pour peu que l'on choisisse:

$$n(N)=N|\log(1-R^2(N))|$$

On en déduit que $Y^N_{n(N)}-Z^*_{n(N)}$ converge en probabilité vers 0.

Par ailleurs, on sait que

lim $\mathrm{loi}(Z^*_{n(N)})=\Phi_{o,\sigma}$ et que

$\mathrm{loi}(Y^N_{n(N)})$=loi stationnaire (M) centrée et normalisée.

On conclut en utilisant le résultat général suivant:

Si deux suites de v.a. X_n et Y_n de f.d.r. respectives F_n et G_n vérifient:

X_n-Y_n tend vers 0 en probabilité.

G_n tend vers G.

alors, F_n tend vers G (cf <Fe71> lemme 2 page254) C.Q.F.D.

4.6. Comparaison asymptotique de l'algorithme SEM et du modèle (M)

Nous allons considérer ici une loi de relance pour l'algorithme SEM induite par le modèle (M).

La différence entre les deux chaines de Markov apparait lorsque la v.a. $T_N(X_n^N)+N^{-1/2}s_N(X_n^N)\eta_{n+1}^N(e;X_n^N)$, que nous noterons $X_{n+1/2}^N$, sort de l'intervalle J_N:

Dans le cas (SEM), on tire X_{n+1}^N selon une loi donnée à l'avance, à support dans J_N.

Dans le cas (M), on pose $X_{n+1}^N=X_{n+1/2}^N$. Mais alors, si $X_{n+1}^N>1-c(N)$, on a

$$X_{n+2}^N=T_N((1-c)+)+N^{-1/2}s_N((1-c)+)\eta_{n+2}^N(e;1-c(N))$$

Si $X_{n+1}^N<c$, la formule est la même en remplacant $(1-c)+$ par $c-$.

Comme $T_N((1-c)+)$ (resp. $T_N(c-)$) est choisi dans $]c(N),1-c(N)[$, $X_{n+2}^N\in J_N$ très probablement:

Il s'introduit ainsi un décalage entre la chaine (SEM) et la chaine (M), d'autant moins important que l'espérance du nombre de sorties de J_N est petite, donc que N est grand. Plus précisément:

<u>Théorème 2:</u>

Notons φ^N la loi stationnaire de la chaine (M). Définissons la chaine (SEM) en choisissant pour loi de relance la probabilité sur J_N induite par l'énoncé suivant:

"Si $X_{n+1/2}^N\notin J_N$, former

$T_N((1-c)+)+N^{-1/2}s_N((1-c)+)\eta_{n+1}^N(e;1-c(N))$

si $X_{n+1/2}^N>1-c(N)$, (le cas où $X_{n+1/2}^N<c(N)$ est analogue)

et prendre X_{n+1}^N égale à cette v.a. si celle-ci est dans J_N.

Sinon, former la suivante en remplacant $\eta_{n+1}^N(e;1-c(N))$ par $\eta_{n+2}^N(e;1-c(N))$,etc."

Notons ψ^N la loi stationnaire de cette chaine (SEM).

Alors, $\varphi^N/J_N < \psi^N$

(le support de ψ^N est inclus dans J_N, au contraire du support de φ^N)

en conséquence, la variation totale de

$|\varphi^N - \psi^N|$ est inférieure ou égale à $\varphi^N(\mathbb{R}-J_N)$,

qui tend vers 0 quand N tend vers l'infini.

Démonstration

Dans cette démonstration, on omet systématiquement l'indice N.

On va suivre une trajectoire de X^{SEM} (noté X) et de X^M, partant du même point intérieur à J, et avec les mêmes entrées $\eta_n(e;p)$:

-tant que X_n^M reste dans J, on a $X_n = X_n^M$.

soit $n=n^*$ l'instant de première sortie de J, i.e. tel que $X_{n+1/2} \notin J$; alors

$X_n^M = X_{n+1/2}$,

tandis que, si l'on tire une v.a. Y_1 selon la loi de relance définie ci-dessus et si $Y_1 \in J$, on a

$X_{n+1} = Y_1$,

auquel cas $X_{n+2}^M = X_{n+1}$;

si $Y_1 \notin J$, on tire, indépendamment, Y_2, et si $Y_2 \in J$ on pose

$X_{n+1} = Y_2 = X_{n+2}^M$, etc.

Ainsi, la suite (X_n^M) passe par les mêmes points de J que la suite (X_n), mais avec un retard croissant. Par ergodicité, cela conduit à notre conclusion. C.Q.F.D.

5. Simulations

Dans toutes les simulations présentées, les algorithmes ont été initialisés en tirant au hasard suivant la loi uniforme les probabilités conditionnelles d'appartenance de chaque point de l'échantillon à l'un des composants du mélange.

Nous présentons les résultats sous forme d'un tableau. Chaque ligne de ce tableau est associée à l'un des paramètres du mélange. La première colonne donne la moyenne des estimés à la stationnarité et la deuxième colonne donne son écart-type. Ces statistiques ont été calculés sur une suite de 100 réalisations. L'écart-type représente l'incertitude relative à l'estimation de chacun des paramètres. Les troisièmes et quatrièmes colonnes donnent les bornes de l'intervalle moyenne-écart-type,moyenne+écart-type. La cinquième colonne donne les valeurs empiriques des paramètres du mélange. Ces valeurs sont bien sûr connues et à découvrir par l'algorithme SEM. Elles sont accompagnés d'une croix lorsque l'intervalle [MOY-E.T.,MOY+E.T.] ne les contient pas.

5.1. mélanges gaussiens unidimensionnels

La taille de l'échantillon est N=200 pour les quatres premiers exemples et N=100 pour le dernier.

Exemple 1

Valeurs théoriques des paramètres:

$K=2, p_1=p_2=0.5, m_1=0, \sigma_1=1, m_2=2, \sigma_2=1$

Initialisation: K=3

Résultats: 2 composants obtenus. Au cours de l'algorithme un composant a disparu. Dans un tel cas, un algorithme de partitionnement ou de reconnaissance de mélange aurait donné le nombre de composants demandés comme d'ailleurs dans tous les exemples présentés.

	MOY.	E.T.	MOY.-E.T.	MOY.+E.T.	VAL.EMP.
p1	0.525	0.084	0.441	0.609	0.500
p2	0.475	0.084	0.391	0.559	0.500
m1	-0.322	0.179	-0.501	-0.143	-0.176
m2	1.968	0.192	1.776	2.160	1.884
s1	1.022	0.167	0.855	1.189	1.042
s2	0.993	0.176	0.817	1.169	1.101

On a noté s la variance de chaque composant.

Exemple 2

Valeurs théoriques des paramètres:

$K=2, p_1=0.25, p_2=0.75, m_1=0, \sigma_1=1, m_2=3, \sigma_2=1$

Initialisation: K=2

Résultats: Ici les proportions sont très différentes, ce qui augmente la difficulté pour une bonne estimation (cf <EvHa81>). L'essai présenté a été initialisé avec le bon nombre de composants, mais d'autres essais montrent que même dans ce cas il suffit de partir d'un majorant de ce nombre de composants.

	MOY.	E.T.	MOY.-E.T.	MOY.+E.T.	VAL.EMP.
p1	0.225	0.032	0.193	0.257	0.250
p2	0.775	0.032	0.743	0.807	0.750
m1	-0.322	0.134	-0.456	-0.188	-0.199
m2	2.808	0.055	2.753	2.863	2.871
s1	0.887	0.170	0.717	1.057	1.008
s2	1.202	0.095	1.107	1.297	1.093

Exemple 3

Valeurs théoriques des paramètres:

$K=2, p_1=0.8, p_2=0.2, m_1=0, \sigma_1=1, m_2=0, \sigma_2=3$

Initialisation: K=3

Résultats: 2 composants obtenus. Dans ce cas les composants sont très imbriqués et de plus les proportions sont très différentes et le nombre de composants initiaux n'est pas le bon.

Naturellement, l'incertitude portant sur les estimations des paramètres est importante.

	MOY.	E.T.	MOY.-E.T.	MOY.+E.T.	VAL.EMP.
p1	0.223	0.055	0.168	0.278	0.200
p2	0.777	0.055	0.722	0.832	0.800
m1	0.086	0.045	0.041	0.131	0.062
m2	0.036	0.187	-0.151	0.223	0.133
s1	0.965	0.114	0.851	1.079	0.990
s2	6.635	1.145	5.490	7.780	6.964

Exemple 4

Valeurs théoriques des paramètres:

$K=1, m=0, \sigma=1$

Initialisation: K=2

Résultats: un seul composant obtenu. Ici il n'y a pas en fait de mélange mais une seule loi normale. l'algorithme SEM permet de découvrir ce fait.

Exemple 5

Valeurs théoriques des paramètres:

$K=4, p_1=p_2=p_3=p_4=0.25$

$m_1=2, \sigma_1=0.5, m_2=5, \sigma_2=0.5, m_3=9, \sigma_3=1, m_4=15, \sigma_4=2$

Cet exemple est tiré de <EvHa81>.

Initialisation: K=5

Résultats: 4 composants obtenus. Notons que quatre valeurs empiriques sont légérement à l'extérieur de l'intervalle MOY-E.T.,MOY+E.T. . Ce fait s'explique par le degré d'imbrication des composants et la taille assez faible des échantillons pour cet exemple (25 par composant).

	MOY.	E.T.	MOY.-E.T.	MOY.+E.T.	VAL.EMP.
p1	0.249	0.003	0.246	0.252	0.250
p2	0.260	0.006	0.254	0.264	0.250(*)
p3	0.250	0.017	0.233	0.267	0.250
p4	0.240	0.014	0.226	0.254	0.250
m1	1.900	0.006	1.888	1.912	1.904
m2	4.937	0.035	4.902	4.972	4.897(*)
m3	9.013	0.040	8.863	9.053	8.835(*)
m4	14.831	0.398	14.433	15.129	14.674
s1	0.161	0.009	0.152	0.170	0.164
s2	0.424	0.063	0.361	0.487	0.340(*)
s3	0.750	0.155	0.595	0.905	0.928(*)
s4	4.476	1.323	3.153	5.799	4.893

Remarque: L'écart-type des estimés des paramètres est le reflet du degré d'imbrication des composants du mélange. Ainsi dans le dernier exemple, les paramètres du premier composant, très éloigné des autres, sont estimés avec une grande précision. Par contre, les estimations des variances des deux composants de l'exemple 3 ont des variances relativement importantes traduisant l'imbrication de ces deux composants au voisinage de l'origine.

5.2. Comparaison des résultats avec ceux obtenus par l'algorihme EM

Dans les exemples 1 à 3 et 5, en partant du bon nombre de composants les résultats sont pratiquement les mêmes pour les deux algorithmes.

Par contre, dans l'exemple 1 en posant K=3, on obtient pour l'algorithme EM:

$p_1=0.322, p_2=0.329, p_3=0.349$
$m_1=0.660, m_2=1.008, m_3=0.887$
$s_1=2.070, s_2=2.128, s_3=2.136$

Et dans l'exemple 4 en posant K=2,on obtient:

$p_1=0.857, p_2=0.143$
$m_1=-1.523, m_2=0.130$
$s_1=0.611, s_2=0.283$

Ainsi l'algorithme EM ne permet pas de déceler l'erreur faite sur le nombre de composants.

En revanche, il donne des résultats plus fiables pour de petits échantillons. Ainsi pour l'exemple 5, si l'on fait passer la taille de l'échantillon de 100 à 60, les résultats par l'algorithme EM ne sont pas notablement modifiés. Mais par l'algorithme d'apprentissage probabiliste, plusieurs essais ont montré qu'environ une fois sur trois l'un des quatre composants disparaissait. Dans les autres cas les résultats sont analogues à ceux obtenus par l'algorithme EM.

Ce fait vient de ce que, pour de petits échantillons, les aléas introduits prennent trop d'importance.

Par contre, il est crucial que la solution initiale de l'algorithme EM soit proche de la solution recherchée.

Ainsi dans l'exemple 1, en ayant posé K=2, nous avons initialisé l'algorithme EM en tirant au hasard les probabilités conditionnelles d'appartenance de chaque point de l'échantillon à l'un des composants du mélange comme nous l'avons fait systématiquement pour l'algorithme SEM. Usuellement, nous avons initialisé l'algorithme EM par l'estimation des paramètres sur la base d'une partition obtenue après une itération de l'algorithme des centres mobiles.

Exemple 1:

A l'itération 1 les valeurs des paramètres étaient (à 10^{-3} près):

$p_1=0.493, p_2=0.507, m_1=0.845, \sigma_1=2.006, m_2=0.862, \sigma_2=2.256$

la valeur de la vraisemblance était de -359.5672

A l'itération 100 les valeurs des paramètres étaient (à 10^{-3} près):

$p_1=0.493, p_2=0.507, m_1=0.840, \sigma_1=2.131, m_2=0.867, \sigma_2=2.134$

la valeur de la vraisemblance était de -359.5159

A l'itération 1000 les valeurs des paramètres étaient (à 10^{-3} près):

$p_1=0.493, p_2=0.507, m_1=0.810, \sigma_1=2.125, m_2=0.896, \sigma_2=2.136$

la valeur de la vraisemblance était de -359.5158

On voit donc qu'au bout de 1000 itérations, les estimations des paramètres (loin des valeurs réelles) n'ont presque pas été modifiées. Dans un tel cas, l'algorithme EM demandera un nombre d'itérations immense pour atteindre des estimations correctes des paramètres.

5.3. Un mélange de Poisson

taille de l'échantillon: N=200

Valeurs théoriques des paramètres:

$K=2, p_1=p_2=0.5, m_1=1, m_2=4$

Initialisation: K=2

Résultats: Cet exemple illustre le fait que l'algorithme SEM peut fonctionner pour tout mélange identifiable.

	MOY.	E.T.	MOY.-E.T.	MOY.+E.T.	VAL.EMP.
p1	0.475	0.055	0.420	0.530	0.500
p2	0.525	0.055	0.470	0.580	0.500
m1	1.126	0.110	1.016	1.236	1.208
m2	3.857	0.130	3.727	3.987	3.911

5.4. Mélanges gaussiens de R^2

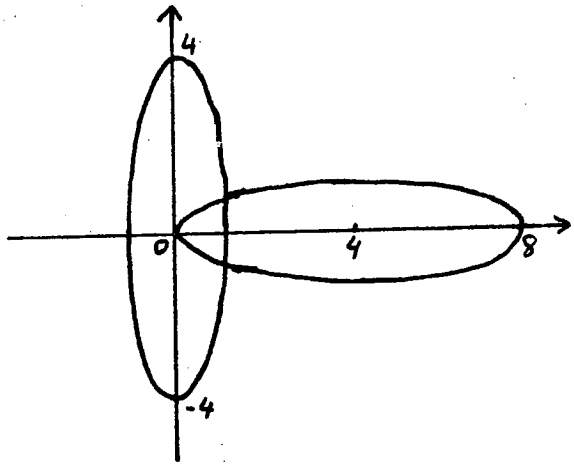
Nous présentons deux exemples d'application sur des données simulées, la taille de l'échantillon étant N=200.

Exemple 1

Valeurs théoriques des paramètres:

$K=2, p_1=p_2=0.5, m_1=(0,0), m_2=(4,0)$

$$\Gamma_1=\begin{bmatrix}1 & 0\\ 0 & 4\end{bmatrix} \quad ; \quad \Gamma_2=\begin{bmatrix}4 & 0\\ 0 & 1\end{bmatrix}$$



ellipsoides d'inertie à 95 %

Initialisation: K=3

Résultats: 2 composants trouvés.

	MOY.	E.T.	MOY.-E.T.	MOY.+E.T.	VAL.EMP.
p1	0.438	0.045	0.393	0.483	0.500(*)
p2	0.562	0.045	0.517	0.607	0.500(*)
m_1^x	-0.253	0.071	-0.324	-0.182	-0.176
m_1^y	-0.110	0.063	-0.173	-0.047	-0.139
Γ_1^x	0.833	0.118	0.715	0.951	1.042(*)
${}_1^{xy}$	-0.303	0.071	-0.374	-0.232	-0.338
Γ_1^y	4.148	0.288	3.860	4.436	3.724(*)
m_2^x	3.407	0.247	3.160	3.654	3.767(*)
m_2^y	0.024	0.045	-0.021	0.069	0.068
Γ_2^x	5.215	0.628	4.587	5.843	4.403(*)
Γ_2^{xy}	-0.068	0.148	-0.080	0.216	-1.166(*)
Γ_2^y	0.974	0.063	0.911	1.037	0.952

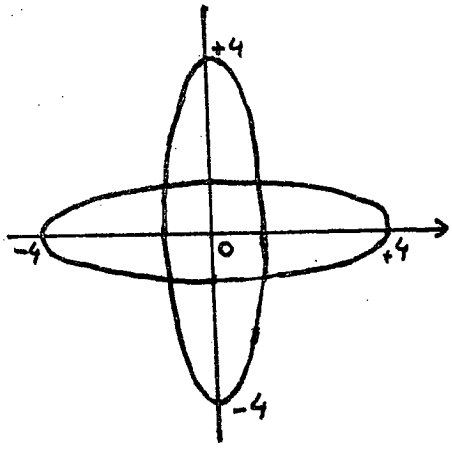
Remarque: l'abscisse de m_2 et ${}_2^x$ sont estimés avec une assez grande imprécision ainsi que les deux proportions.

Exemple 2

Valeurs théoriques des paramètres:

$K=2, p_1=p_2=0.5, m_1=(0,0), m_2=(0,0)$

$$\Gamma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \quad ; \quad \Gamma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



ellipsoides d'inertie à 95 %

Initialisation: K=2

Résultats: Les moyennes sont identiques, malgré cela les paramètres sont correctement estimés. Du point de vue clasification par contre, il n'est pas possible de distinguer deux classes.

	MOY.	E.T.	MOY.-E.T.	MOY.+E.T.	VAL.EMP.
p1	0.526	0.071	0.455	0.597	0.500
p2	0.474	0.071	0.403	0.545	0.500
m_1^x	0.110	0.071	0.039	0.181	0.089
m_1^y	0.103	0.076	0.026	0.180	0.117
Γ_1^x	0.868	0.138	0.730	1.006	0.823
Γ_1^{xy}	0.221	0.084	0.137	0.305	0.020(*)
Γ_1^y	4.362	0.365	3.999	4.725	4.575
m_2^x	-0.186	0.095	-0.281	-0.091	-0.212
m_2^y	0.000	0.084	-0.084	0.084	-0.022
Γ_2^x	4.572	0.460	4.112	5.032	4.405
Γ_2^{xy}	-0.301	0.114	-0.415	-0.187	-0.067(*)
Γ_2^y	1.015	0.232	0.783	1.247	0.967

5.5. Un mélange gaussien dans R^5

Nous présentons un exemple dans un espace de dimension supérieure à deux sur des données simulées, la taille de l'échantillon étant N=400.

Valeurs théoriques des paramètres:

$K=3, p_1=0.5, p_2=p_3=0.25$

$m_1=(0,1,0,0,0)$; $m_2=(0,0,3,0,0)$; $m_3=(0,0,0,4,0)$

$$\Gamma 1=\begin{bmatrix}1&0&0&0&0\\0&2&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix};\quad \Gamma 2=\begin{bmatrix}4&0&0&0&0\\0&4&0&0&0\\0&0&1&0&0\\0&0&0&3&0\\0&0&0&0&4\end{bmatrix};\quad \Gamma 3=\begin{bmatrix}16&0&0&0&0\\0&16&0&0&0\\0&0&4&0&0\\0&0&0&1&0\\0&0&0&0&16\end{bmatrix}$$

Valeurs empiriques des paramètres $(a_k, k=1,K)$:

$m_1=(0.032,0.962,0.049,0.066,-0.108)$

$m_2=(-0.016,-0.122,2.923,0.012,-0.175)$

$m_3=(0.272,-0.290,0.197,3.787,0.143)$

$$\Gamma 1=\begin{bmatrix} 0.827 & 0.020 & 0.127 & 0.087 & -0.069 \\ 0.020 & 1.779 & 0.085 & -0.082 & 0.113 \\ 0.127 & 0.085 & 1.136 & 0.102 & -0.087 \\ 0.087 & -0.082 & 0.102 & 0.970 & -0.084 \\ -0.069 & 0.113 & -0.087 & -0.084 & 0.948 \end{bmatrix}$$

$$\Gamma 2=\begin{bmatrix} 3.990 & 0.560 & -0.057 & 0.488 & -0.350 \\ 0.560 & 3.660 & 0.494 & 0.236 & -0.724 \\ -0.057 & 0.494 & 1.062 & 0.226 & -0.242 \\ 0.488 & 0.236 & 0.226 & 2.871 & -0.512 \\ -0.350 & -0.724 & -0.242 & -0.512 & 3.413 \end{bmatrix}$$

$$\Gamma 3=\begin{bmatrix} 16.850 & 0.594 & 0.474 & 0.093 & 1.085 \\ 0.594 & 17.818 & 1.252 & -0.032 & 0.963 \\ 0.474 & 1.252 & 3.514 & 0.463 & -0.244 \\ 0.093 & -0.032 & 0.463 & 1.088 & -0.462 \\ 1.085 & 0.963 & -0.244 & -0.462 & 18.444 \end{bmatrix}$$

Il faut noter que les valeurs empiriques des paramètres sont sensiblement différentes des valeurs théoriques.

Initialisation: K=6

Nous partons d'un nombre de composants nettement supérieur au nombre réel de composants. Cet exemple illustre bien la capacité pour l'algorithme SEM de se fixer sur le bon nombre de composants en partant d'un majorant de ce nombre.

En effet, à la convergence, l'algorithme SEM se fixe sur trois composants que nous présentons ci-après.

Résultats:

Afin de ne pas alourdir le texte, nous ne donnons que la moyenne de la suite des estimés à la stationnarité.

$p_1=0.508$, $p_2=0.245$, $p_3=0.247$

$m_1=(0.035,0.886,0.087,0.085,-0.157)$

$m_2=(0.021,0.074,2.947,-0.027,-0.144)$

$m_3=(0.232,-0.361,0.153,3.822,0.214)$

$$\Gamma 1=\begin{bmatrix} 0.772 & 0.060 & 0.145 & 0.012 & -0.032 \\ 0.060 & 1.802 & -0.028 & -0.115 & 0.206 \\ 0.145 & -0.028 & 1.212 & 0.128 & -0.154 \\ 0.012 & -0.115 & 0.128 & 0.950 & -0.074 \\ -0.032 & 0.206 & -0.154 & -0.074 & 0.869 \end{bmatrix}$$

$$\Gamma 2=\begin{bmatrix} 4.205 & 0.582 & -0.021 & 0.720 & -0.491 \\ 0.582 & 3.997 & 0.280 & 0.491 & -0.551 \\ -0.021 & 0.280 & 1.125 & 0.363 & -0.538 \\ 0.720 & 0.491 & 0.363 & 2.831 & -0.609 \\ -0.491 & -0.551 & -0.538 & -0.609 & 3.368 \end{bmatrix}$$

$$\Gamma 3=\begin{bmatrix} 16.946 & 0.506 & 0.306 & 0.171 & 0.158 \\ 0.506 & 17.807 & 1.068 & 0.033 & 0.741 \\ 0.306 & 1.068 & 3.329 & 0.549 & 0.100 \\ 0.171 & 0.033 & 0.549 & 1.044 & -0.655 \\ 1.158 & 0.741 & 0.100 & -0.655 & 18.792 \end{bmatrix}$$

Il apparait que les estimations des paramètres sont très proches des valeurs empiriques. Signalons, d'autre part, que les variances de ces estimés sont faibles.

Enfin, du point de vue classification, le croisement de la partition $P=(P_1,P_2,P_3)$ obtenue à l'itération qui fournit la plus grande valeur de la vraisemblance avec la partition $O=(O_1,O_2,O_3)$ d'origine donne le tableau de confusion suivant:

	O1	O2	O3	
P1	191	11	1	203
P2	9	86	3	98
P3	0	3	96	99
	200	100	100	400

Il y a 27 points mal classés sur 400, soit 6,75%. Ces résultats sont satisfaisants, compte-tenu du fait que les zones de recrouvrement des trois composants ne sont pas négligeables, et témoignent de la bonne qualité des estimations.

5.6. Evaluation de la qualité de l'approximation du théorème 1

Le théorème de convergence de 4.5 considére une modélisation approximative de l'algorithme SEM valable lorsque N tend vers l'infini.

Il est important de voir si à N fixé, l'approximation de la loi $N^{1/2}(p_S^N-p_N)$ par une loi normale centrée et de variance

$\sigma^2=s^2(p)(1-\rho^2)^{-1}$

est raisonnable. Pour ce faire, nous avons repris l'exemple 2 du paragraphe 5.1.1 (N=200) en considérant seul p=0.25 comme inconnu.

A la stationnarité (après 100 réalisations de la loi stationnaire) on obtient:

$MOY(p_N)=0.246$ $E.T.(p_N)=0.137$

D'autre part une estimation de p_N par l'algorithme EM donne $p_N=0.247$.

d'où $N^{1/2}(p_S^N-p_N)$ a pour moyenne -0.014 et pour écart-type 0.1937.

Par ailleurs, une estimation de σ sur la base de l'échantillon donne $\sigma=0.207$.

Enfin le tableau suivant (résultat d'un histogramme à intervalles égaux) permet de juger de l'approximation de

$N^{1/2}(p_S^N-p_N)$

par une loi normale centrée d'écart-type 0.207.

classe	minimum	maximum	freq.obs.	freq.theo.
no 1	-infini	-0.450	0.025	0.015
no 2	-0.450	-0.307	0.045	0.056
no 3	-0.307	-0.164	0.190	0.145
no 4	-0.164	-0.028	0.255	0.231
no 5	-0.028	0.121	0.245	0.273
no 6	0.121	0.264	0.190	0.178
no 7	0.264	0.407	0.050	0.077
no 8	0.407	+infini	0.000	0.025

La statistique du CHI2 est égale à 6.34. p_N et σ ayant été estimés, cette valeur est à comparer à un CHI2 à 5 degrés de liberté.

Au seuil de 0.1 la valeur critique du CHI2 est de 9.24.

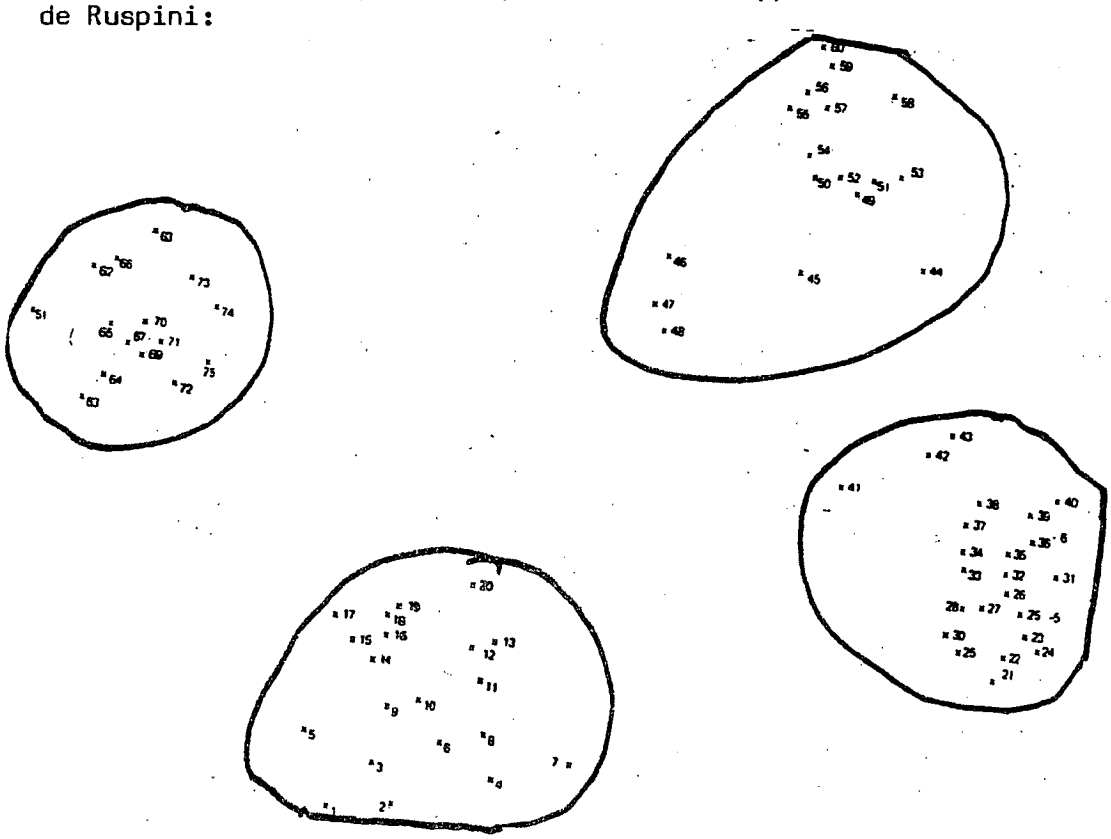
Au vu de ces résultats sur cet exemple, l'approximation semble très bonne.

6. Applications à des données réelles

6.1. Données de Ruspini

Bien entendu, comme nous l'avons expliqué aux paragraphes 1 et 2, l'application de cette méthode dépasse le cadre de reconnaissance de mélanges et peut être vue comme une méthode de classification permettant en particulier de s'affranchir de la connaissance du nombre de classes.

A titre d'exemple nous présentons son application sur les données de Ruspini:



Données de Ruspini

Initialisé avec K=5, l'algorithme donne bien les 4 classes naturelles entourés sur la figure. Notons que la variance des estimations des paramètres est nulle car les classes sont bien séparées.

6.2. Application à des données météorologiques

Les données de cette application nous ont été fournies par G.Seze et C.Belcour (laboratoire de météorologie dynamique du CNRS).

Il s'agit de données concernant la couverture nuageuse au dessus du Sud-Ouest de la France et de l'Espagne. Un satellite mesure la lueur visible et la lueur infra-rouge réfléchie par les nuages.

Le problème est de distinguer les nuages bas, les nuages moyens et les nuages hauts, les frontières entre ces trois types de nuages étant assez floues.

Nous avons travaillé sur trois jours assez bien différenciés. Le premier jour correspondait plutôt à des nuages bas, le deuxième plutôt à des nuages moyens et le troisième plutôt à des nuages hauts. Les données pour chaque jour sont constituées de 2500 points sur lesquels sont mesurés les deux paramètres.

Nous disposions donc de 7500 points décrits par deux paramètres.

Nous avons soumis ces données à l'algorithme SEM avec une modélisation Gaussienne en demandant 6 composants.

A la convergence les 6 composants obtenus s'interprêtent comme:

- le sol découvert
- la mer découverte
- les nuages bas
- les nuages moyens
- les nuages hauts
- des nuages intermédiaires entre bas et moyens.

Ces résultats ont été comparés avec ceux obtenus par les nuées dynamiques en travaillant sur les données centrées réduites.

D'après G.Seze et C.Belcour, il ressort que les centres de gravité des classes obtenus dans les deux cas sont assez similaires. Tout au plus, ceux obtenus par l'algorithme SEM fournissent des estimations légérement plus précises pour les centres de gravité des trois zones nuages bas, nuages moyens, nuages hauts.

Par contre, les composants correspondant à ces trois zones obtenus par l'algorithme SEM les cernent beaucoup mieux que les classes correspondantes obtenues par les nuées dynamiques. L'information apportée par les composants de l'algorithme SEM s'avère plus riche du point de vue des météorologues.

Ce résultat s'explique sans doute par le fait que les nuées dynamiques avec la métrique euclidienne usuelle ont tendance à donner des classes sphériques comme on l'a montré au paragraphe 2.1.3.

L'intérêt présenté par les résultats obtenus par l'algorithme SEM est d'autant plus remarquable que la modélisation gaussienne retenue pour des raisons numériques était, de l'avis même de G.Seze et C.Belcour, une approximation grossière de l'idée physique qu'elles se faisaient a priori de la forme des lois des composants.

Cette application fait ressortir que l'algorithme SEM peut être particulièrement utile en classification dans le cas où les classes à "découvrir" ont des frontières floues, pour peu que la modélisation choisie soit raisonnable.

7. Version séquentielle stricte de l'algorithme SEM

Dans son principe, la version séquentielle stricte de l'algorihme SEM consiste à considérer que l'échantillon est construit par étapes successives. A chaque étape n, on observe le point x_n et l'on modifie les estimations des paramètres du mélange compte tenu de ce nouveau point.

Nous parlons ici de version séquentielle stricte car, comme on le voit ci dessous, seul le dernier point observé est pris en compte pour la modification des estimations. Comme on le verra au paragraphe 8, une autre version séquentielle (large) est possible. Elle consiste à tenir compte de tous les points déjà observés $(x_1,\ldots,x_n)$ pour la modification des paramètres.

Du point de vue pratique, le fait de modifier les estimations des paramètres en ne tenant plus compte, à chaque itération, de tout l'échantillon mais seulement d'un point ralentit considérablement la convergence vers la loi stationnaire.

Par contre, le cadre séquentiel peut être profitable pour étudier le comportement asymptotique de l'algorithme car nous disposons alors des théorèmes de convergence de martingales.

7.1. Présentation

Initialisation

Le nombre K de composants étant fixé, on choisit (en général au hasard) les valeurs initiales

$(q_k^o)=(p_k^o,a_k^o);k=1,K)$

des paramètres.

Itération n: $(n>0)$

Soit x_n le nouveau point considéré de l'échantillon.

Etape S

On tire en ce point la v.a. multinomiale $e^n(x_n)=(e_1^n(x_n),\ldots,e_K^n(x_n))$ d'ordre 1 et de paramètre

$(t_1^n(x_n),\ldots,t_K^n(x_n))$

Etape M

Modification des estimations des paramètres du mélange compte tenu du tirage $e_n(x_n)$:

$\forall k=1,K\ p_k^n=((n-1)p_k^{n-1}+e_k^n(x_n))/n$

La modification des estimations des paramètres $(a_k,k=1,K)$ dépend de la famille paramétrée des composants du mélange. Dans le cas courant où l'espérance m_k et la matrice variance Γ_k sont des constituants des paramètres, il vient pour k=1,K:

$m_k^n=((n-1)p_k^n m_k^{n-1}+e_k^n(x_n)x_n)/((n-1)p_k^{n-1}+e_k^n(x_n))$

et

$\Gamma_k^n=((n-1)p_k^{n-1}\ {}_k^{n-1}$

$+(n-1)p_k^{n-1}(m_k^{n-1}-m_k^n)(m_k^{n-1}-m_k^n)'+e_k^n(x_n-m_k^n)(x_n-m_k^n)')/((n-1)p_k^{n-1}+e_k^n(x_n))$

Etape E

Pour k=1,K calcul de:

$t_k^{n+1}(x_{n+1})=p_k^n f(x_{n+1},a_k^n)/\ \Sigma(p_k^n f(x_{n+1},a_k^n),k=1,K)$

7.2. Comportement asymptotique dans le cas d'un mélange à deux composants, où seules les proportions sont inconnues

On peut se ramener à l'estimation du seul paramètre $p=p_1$ du fait que $p_1+p_2=1$.

L'algorithme se simplifie de la manière suivante:

Initialisation

Le nombre K de composants étant fixé, on choisit au hasard p^o valeur initiale de p.

Itération n: (n>0)

Soit x_n le nouveau point considéré de l'échantillon.

Etape S

On tire en ce point la v.a. de Bernoulli $e^n(x_n)$ de paramètre $t^n(x_n)$.

Etape M

Modification de l'estimation de p compte tenu du tirage $e_n(x_n)$:

$$p^n=((n-1)p^{n-1}+e^n(x_n))/n$$

Etape E

Calcul de:

$$t^{n+1}(x_{n+1})=p^n f(x_{n+1},a_1)/(p^n f(x_{n+1},a_1)+(1-p^n)f(x_{n+1},a_2)$$

<u>Proposition:</u>

Dans le cadre défini ci dessus, la suite des p_N de l'algorithme SEM séquentiel strict converge presque sûrement vers la vraie valeur p* de la proportion à estimer.

Démonstration:

Nous ne présentons pas en détail la démonstration de cette proposition. Elle est, en effet, quasiment identique à celle de Silverman <Si80> pour son algorithme d'apprentissage probabiliste dans un cadre Bayesien (cf paragraphe 2.2.4).

Nous nous contenterons d'en indiquer les grandes lignes en explicitant certains points non détaillés par Silverman.

Notons:

$$T(p)= \int pf(x,a_1)f^*(x)/pf(x,a_1)+(1-p)f(x,a_2))dx$$

et $S(p)=T(p)-p$

Soit F_n la tribu engendrée par $(p_1,\ldots,p_n)$. On montre que:

$$E((p_{n+1}-p^*)^2/F_n)=(p_n-p^*)^2+2(p_n-p^*)S(p_n)/(n+1)$$

$+((1-p_n)^2T(p_n)+p_n^2(1-T(p_n)))/(n+1)^2$

Soit la suite de v.a.

$Z_n=(p_n-p^*)^2-\Sigma(1/r^2,r=1,n)$.

On déduit de l'égalité précédente que:

$E((p_{n+1}-p^*)^2/F_n)\leqslant(p_n-p^*)^2+1/(n+1)^2$

d'où $E(Z_{n+1}/F_n)\leqslant Z_n$

Z_n est donc une sur-martingale bornée et il existe une v.a. Z telle que Z_n converge vers Z p.s.

Il s'en suit qu'il existe une v.a. p_+ telle que p_n converge vers p_+ p.s.

En effet, $\forall\omega\in\Omega_o$, avec $P(\Omega_o)=1$, $(p_n-p^*)^2(\omega)$ converge vers $Z(\omega)$.

Donc la suite $p_n(\omega)$ a au maximum deux valeurs d'adhérence

$l(\omega)=p^*+\sqrt{Z}(\omega)$, $l'(\omega)=p^*-\sqrt{Z}(\omega)$ et en a au moins une.

Or, il est facile de voir que $|p_{n+1}-p_n|\leqslant 1/(n+1)$.

On en déduit que pour tout $\varepsilon>0$ il existe n (suffisamment grand) tel que:

$|p_n(\omega)-l(\omega)|<\varepsilon$, $|p_{n+1}(\omega)-l'(\omega)|<\varepsilon$, $|p_{n+1}(\omega)-p_n(\omega)|<\varepsilon$

d'où:

$|l(\omega)-l'(\omega)|<|p_n(\omega)-l(\omega)|+|p_{n+1}(\omega)-p_n(\omega)|+|p_{n+1}(\omega)-l'(\omega)|<3\varepsilon$

On a donc nécessairement $l(\omega)=l'(\omega)$. La suite $p_n(\omega)$ définie sur le compact $[0,1]$ a une seule valeur d'adhérence et donc converge vers cette valeur. Donc pour tout ω de Ω_o :

$p_n(\omega)$ converge vers une fonction de ω de la forme

$p^*+\delta(\omega)\sqrt{Z}(\omega)$ avec $\delta(\omega)=+1$ ou -1.

Cette fonction limite de v.a. est une v.a.

Silverman montre ensuite que $(p^*-p_+)S(p_+)=0$ ce qui entraine que p_+ ne peut être que 0, 1 ou p^*.

Soit maintenant l'événement

$A_n=\{|p_n-p^*|<1/n\}$.

Supposant, par commodité, que $0<p_o<p^*$ Silverman montre que p_n^t, i.e. le processus p_n arrêté au premier pas t pour lequel A_n est réalisé est minoré par le processus

$$\bar{p}_{n+1}=(n\bar{p}_n+e_{n+1}(x_{n+1})b_n)/(n+1)$$

b_n étant une v.a. de Bernoulli indépendante des autres v.a. avec

$$P(b_n=1)=\bar{p}_n/T(p_n)<1$$

Or ce processus est une martingale associée à une urne de Polya et donc la probabilité que $\bar{p}_n$ converge vers 0 est nulle (cf <Fe71> page 243). D'où

$$\lim p_n^t>p^* \text{ p.s.}$$

Donc soit t est plus petit que l'infini (auquel cas A_n a lieu au moins une fois) soit $p+=p^*$.

Montrons maintenant que pour tout m, on a p.s.:

Soit p_n converge vers p^*.

Soit il existe $n>m$ tel que A_n a lieu.

En effet, si $t=\infty$ p_n converge vers p^*. Sinon:

Soit m quelconque,

-Si $p_m<p^*$ on est ramené au raisonnement précédent et donc il existe $n>m$ tel que A_n a lieu.

-Si $p_m\geqslant p^*$, considérons le processus $\hat{p}_n=p_{n+m}$

On a $p^*\leqslant\hat{p}_o\leqslant 1$.

Soit s le premier pas tel que $\hat{p}_s<p^*$.

Considérons $\hat{p}_n^s$ le processus $\hat{p}_n$ stoppé à l'instant s.

De manière analogue à précédemment, on peut montrer que le processus $\hat{p}_n^s$ est majoré par le processus $\tilde{p}_n$ défini par:

$$1-\tilde{p}_{n+1}=(1-\tilde{p}_n+1-e_{n+1}(x_{n+1})c_n)/(n+1)$$

c_n étant une v.a. de Bernoulli indépendante des autres variables avec:

$$P(c_n=1)=(1-\tilde{p}_n)/T(p_n)<1.$$

Le processus $(1-\tilde{p}_n)$ est une martingale associée à une urne de Polya et donc la probabilité que $\tilde{p}_n$ converge vers 1 est nulle.

D'où $\lim \hat{p}_n^s < p^*$ et donc:

Soit s est plus petit que l'infini auquel cas A_n a lieu.

Soit $p+=p^*$.

Finalement on a p.s.:

Soit $p+=p^*$.

Soit A_n a lieu un nombre infini de fois.

Mais si $p+\neq p^*$, A_n a lieu un nombre fini de fois.

Donc $p+=p^*$. C.Q.F.D.

Remarque

Pour son algorithme, on a vu que Silverman se placait dans le cadre bayésien et partait d'une distribution bêta pour p_0 dont il incrémentait l'un ou l'autre des paramètres suivant le résultat des tirages aléatoires (cf paragraphe 2.2.4).

On voit que ce formalisme n'introduit que des différences techniques minimes avec l'algorithme SEM séquentiel strict. Ces deux algorithmes sont équivalents tant du point de vue pratique que du point de vue du comportement asymptotique.

8. Version séquentielle large de l'algorithme

8.1. Présentation

Cette version différe de la précédente dans le fait qu'à chaque nouvelle observation x_{n+1}, on réeffectue l'ensemble des opérations de l'algorithme sur la base de l'échantillon déjà observé $(x_1,\ldots,x_{n+1})$.

Nous allons étudier le comportement asymptotique de cet algorithme (dans le cas où seul p* est inconnu) en utilisant une modélisation analogue à celle du paragraphe 4.2. L'étude qui suit s'appuie donc sur les résultats obtenus en 4. La modélisation devient:

(MSL) $X_{n+1}=T_n(X_n)+n^{-1/2}(X_n)\eta_{n+1}(e;X_n)$

avec une formule analogue pour $\bar{X}_n=X_n-p_n$, en notations centrées.

8.2. Etude du comportement asymptotique

la chaine (MSL) est non homogène, et la question qui se pose est celle de la convergence en probabilité de la suite des v.a. X_n vers p*, lorsque n tend vers l'infini.

Pour préciser la vitesse de cette convergence, on introduit la suite des v.a.

$Y_n=n^{1/2}\bar{X}_n$,

et on montre que, quand n tend vers l'infini, la loi de Y_n tend vers une loi normale centrée.

Il reste à montrer que si l'on choisit, pour l'algorithme SEM séquentiel large, une loi de relance induite par le modèle (MSL), si l'on note V_n la suite des v.a. ainsi engendrée, si l'on forme

$W_n=n^{1/2}\bar{V}_n$,

alors la loi de W_n converge vers une loi normale centrée. Ceci est fait en supposant:

$T_n((1-c)+)=\text{Cste} \in \,]c(n),1-c(n)[$

$T_n(c-)=\text{Cste} \in \,]c(n),1-c(n)[$ et $s_n(p)=0$ hors de $J(n)=[c(n),1-c(n)]$.

Il en résultera, que dans ce cas séquentiel large, la suite $(V_n;n>0)$ converge en probabilité vers p^* lorsque n tend vers l'infini et de plus la loi de V_n suit approximativement une loi normale centrée sur p^*.

Théorème 3:

Si $c(n)$ tend vers 0 assez lentement, si l'on note X_n les v.a. définies par le modèle (MSL), si l'on note

$Y_n=n^{1/2}(X_n-p_n)$, alors:

(i) X_n tend vers p^* en probabilité quand n tend vers l'infini.

(ii) $P(Y_n<a)$ tend vers $\Phi_{0,\sigma}(a)$ quand n tend vers l'infini, avec les notations du théorème 1 et:

$\sigma^2=s^{*2}(1-\rho^{*2})^{-1}$.

Démonstration:

On reprend point par point celle du théorème 1 (cf 4.5.). On a comme précédemment:

$E(Y_n)^2<(1-R(n)^2)^{-1}$

$E(Y_n)^4<Cste(1-R(n))^{-4}$

$|n^{1/2}\bar{T}_n(qn^{-1/2})-\rho_n q|<C_1 q^2 n^{-1/2}B(n)$

$|\bar{s}_n(qn^{-1/2})-\bar{s}_n(0)|<C_2|q|n^{-1/2}G(n)$

$Y_{n+1}(n/n+1)^{1/2}=\rho_n Y_n+A(Y_n)+\bar{s}_n(0)\eta_{n+1}(e,Y_n n^{-1/2})+D(Y_n)\eta_{n+1}(e,Y_n n^{-1/2})$

C'est-à-dire, Y_{n+1} à la même forme qu'en 4.5 où on multiple $A(Y_n)$ et $D(Y_n)$ par:

$(n/n+1)^{1/2}=1-1/2n+o(1/n)$.

On pose encore $u_n=E^{1/2}|Y_n-Z_n|^2$ et on trouve:

$u_{n+1}\leqslant\rho_n u_n+r(n)$ et $u_o=0$

avec $\lim r(n)=0$ et $\lim \rho_n=\rho^*$.

Soit $U_{n+1}=\rho_n U_n+r(n)$ avec $v_o=0$; par récurrence, on trouve:

$u_n\leqslant U_n$.

$U_n=\Sigma(\rho_{n-1}\cdots\rho_{n-j-1}r(j);j=0,n-1)$

qui tend vers 0 quand n tend vers l'infini. Puis:

$E^{1/2}|Y_n-Z^*_n|^2<\rho^{*n}(1-R^2(n))^{-1/2}+V(n)$ avec $\lim V(n)=0$.

tend vers 0 quand n tend vers l'infini, si c(n) tend vers 0 assez lentement.

En effet, le log de $\rho^{*n}(1-R^2(n))^{-1/2}$ s'écrit:

$n\log\rho^*+1/2|\log(1-R^2(n))|$

Cette quantité tend vers moins l'infini dès que

$|\log(1-R^2(n))|=o(n)$

Ceci est obtenu si c(n) tend vers 0 assez lentement.

Ainsi $|P(Y_n<a)-P(Z^*_n<a)|$ tend vers 0 pour tout a.

Soit $|P(Y_n<a)-\Phi_{0,\sigma}(a)|$ tend vers 0 pour tout a.

avec $\sigma^2=s^{*2}(1-\rho^{*2})^{-1}$.

Ceci implique que $P(|\bar{X}_n|>\varepsilon)$ tend vers 0 (quand n tend vers l'infini) pour tout $\varepsilon>0$, et, plus précisément, que:

$P(|\bar{X}_n|>an^{-1/2})$ tend vers $2(1-\Phi_{0,\sigma}(a))$, pour $a>0$.

Enfin, comme il a été fait en 4.6, la propositon suivante montre que l'écart entre le modèle (MSL) et l'algorithme SEM devient négligeable quand n tend vers l'infini.

<u>Théorème 4:</u>

Supposons de plus que $s_n(p)=0$ hors de J(n). Alors, la loi limite, quand n tend vers l'infini, de

$n^{1/2}\bar{X}_n$

est la même que celle de

$n^{1/2}\bar{X}^M_n$.

Démonstration:

Définissons U(n) récursivement par:

-U(n)=0, si $X_n=X^M_n$ est resté dans J(n).

-U(n)=1 si $X^M_{n+1/2}\notin J(n)$ pour la première fois;

jusqu'à la sortie suivante on gardera U(n)=1, et on aura l'égalité:

$X^M_{n+1}=X_n$ si $X^M_{n+1}\epsilon J(n)$.

$-U(n)=m$ entre l'instant de m^{ieme} sortie de X_n et l'instant de $(m+1)^{ieme}$ sortie.

On a alors $X^M_{n+U(n)}=X_n$ lorsque $X^M_{n+U(n)}\epsilon J(n)$.

Puisque le bruit additif a une variance finie qui tend vers 0 quand n tend vers l'infini, $U(n)/n$ tend en probabilité vers 0. En conséquence:

$$\text{loi}(\bar{X}^M_{n+U(n)}(n+U(n))^{1/2})=\text{loi}(\bar{X}_n n^{1/2}(1+n^{-1}U(n))^{1/2}).$$

BIBLIOGRAPHIE

<Ag70> AGRAWALA-"Learning with a probabilistic teacher" IEEE.Information theory,Vol 16,nu.4.

<Bi68> BILLINGSLEY-"Convergence of probability measures" Wiley.

<Bo83> BOYLES-"On the convergence of the EM algorithm" JRSS.B.Vol 45,nu.1.

<Br68> BREIMAN-"Probability" Addison-Wesley.

<BCD81> BRONIATOWSKI,CELEUX,DIEBOLT-"Clustering by gaussian mixture recognition methods" Proc.of sixth symposium on Operational Research.

<BCD83> BRONIATOWSKI,CELEUX,DIEBOLT-"Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste" Actes des troisièmes journées internationales d'analyse des données Versailles Oct 83. Ed: North Holland.

<BrWy78> BRYANT,WYLLIAMSON-"Asymptotic behaviour of classification maximum likelihood estimates" Biometrika 78,Vol 68.

<Ca76> CAZES-"Décomposition d'un histogramme en composantes gaussiennes" Revue de statistique appliquée Vol 24 No 4.

<Co64> COOPER-"Non supervised adaptative signal detection and pattern recognition" Information and Control Vol 7.

<Da69> DAY-"Estimating the components of a mixture of normal distributions" Biometrika 69,Vol 56.

<Di80> DIDAY et collaborateurs-"Optimisation en classification automatique" Editeur: INRIA

<DLR77> DEMPSTER,LAIRD,RUBIN-"Maximum likelihood from incomplete data via the EM algorithm" JRSS.b. Vol 39.

<EvHa81> EVERITT,HAND-"Finite mixture distributions" Chapman and Hall.

<Fe71> FELLER-"An introduction to probability theory and its applications" Vol2,2.ed.,Wiley.

<Go75> GOVAERT-"Classification automatique et distances adaptatives" Thèse de troisième cycle Université Paris 6 1975.

<Ho78> HOSMER-"Comment on Quandt and Ramsey paper" JASA Vol 73.

<Ka77> KAZAKOS-"Recursive estimation of prior probabilities using a mixture" IEEE.Information theory,Vol 23.

<KeSn74> KEMENY,SNELL-"Finite Markov chains" Springer-Verlag.

<KeSt73> KENDALL,STUART-"The advanced theory of statistics" Vol2,3.ed.,Griffin.

<Ki78> KIEFER-"Efficient estimation of a switching regression model" Econometrica,Vol 46,nu 2.

<Ma75> MARRIOTT-"Separating mixtures of normal distributions" Biometrika 31.

<MaSm76> MAKOV,SMITH-"Quasi Bayes procedures for unsupervised learning" Proc IEEE.Conf. on Decision and Control.

<Pe94> PEARSON-"Contribution to the mathematic theory of evolution" Philos.Trans.Soc..nu 185 (1894).

<QuRa78> QUANDT,RAMSEY-"Estimating mixtures of normal distributions and switching regression" JASA Vol 73.

<ReWa84> REDNER,WALKER-"Mixture densities,maximum likelihood and the EM algorithm" SIAM Review Vol 26 No 2 April 84.

<Re75> REVUZ-"Markov chains" North Holland.

<Sch76> SCHROEDER-"Analyse d'un mélange de distribution de probabilité de même type" RSA,Vol 24,nu 1.

<ScSy71> SCOTT,SYMONS-"Clustering methods based on likelihood ratio criteria" Biometrics Vol 27.

<Sh68> SHLEZINGER-"An algorithm for solving the selforganization problem" Cybernetics nu 2.

<Si80> SILVERMAN-"Some asymptotic properties of the probabilistic teacher" IEEE.Information theory,Vol 26,nu 2.

<Sk56> SKOROHOD-"Limit theorems for stochastic processes" Theory of probability and its applications nu 1.

<SmMa78> SMITH,MAKOV-"A quasi Bayes sequential procedure for mixtures" JRSS.B.Vol 40,nu 1.

<Te62> TEICHER-"Identifiability of finite mixture" Ann.Math.Statist. Vol 34.

<Sy81> SYMONS-"Clustering criteria and multivariate normal mixtures" Biometrics Vol 37.

<Wo70> WOLFE-"Pattern clustering by multivariate mixture analysis" Multiv.Behav.Res. Vol 5.

<Wu83> WU-"On the convergence of the EM algorithm" Ann.Statist. 83 Vol 11 No 1.

<YaSp68> YAKOWITZ,SPRAGINS-"On the identifiability of finite mixtures" Ann. Math. Statist., Vol 39.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique